CLRES 2707
Bioinformatics Resources: Data Mining

CLRES 2708
Bioinformatics Resources: Data Analysis

Course Instructor(s):	Email address(es)
Ansuman Chattopadhyay, PhD	ansuman@pitt.edu
Carrie Iwema, PhD, MLS	iwema@pitt.edu
Uma Chandran, PhD, MSIS	chandran@pitt.edu

Dates:
Data Mining: 2/21/13 - 3/26/13
(no class on 3/12 and 3/14)

Data Analysis: 3/28/13 - 4/23/13

Meeting days, time:
T TH 10:00 -12:00

Location:
Falk Library Classroom 2

Phone contact:
412-648-1297 (Chattopadhyay)
412-383-6887 (Iwema)

Registration permission number
Contact: Jennifer M. Holloman
Email: hollimanjm@upmc.edu
412-586-9673

Overview and Objectives:

Over the past decade, the emergence and rapid advances in molecular technologies such as genome sequencing, microarray platforms, and high-throughput methodologies have generated a copious amount of scientific data. In response to this data overload, bioinformatics software and databases utilizing computer science and statistical methods have rapidly evolved. Proficiency in the use of bioinformatics tools is the key for success in today's molecular life sciences research as they guide students in formulating new hypotheses, designing studies to test these hypotheses, then analyzing, interpreting and validating experimental results.

This course is divided into **two 1-credit hour classes** and will be offered in a computer classroom to provide hands-on training experience. **CLRES 2707** focuses on **data mining** while **CLRES 2708** provides training on **data analysis**. Upon successful completion of this course, students will have received adequate training to identify appropriate bioinformatics databases/software and efficiently apply these tools in solving real life research questions.

Students are encouraged to register for both classes in the same semester to receive maximum benefit in terms of class continuity.

Responsibilities:

- Reading assignments are strongly recommended to be completed before each class.

- Students must complete homework assignments by set due dates. Multimedia video presentations using screen capturing software for class lecture materials will be available to assist with solving homework questions. Students are encouraged to work together on class projects and homework, but should write up results individually. Homework assignments are to be turned in via email by 11 pm on the due date.

- **CLRES 2707**: students must prepare a gene report to be turned in on **March 28**. Students may use their own gene of interest, or an instructor-assigned gene if needed, to write a report highlighting gene and protein based information gathered from the various databases introduced in this course.

- **CLRES2708**: students must prepare a research proposal on their gene (from first class) to be (1) presented as a short PowerPoint presentation in the final class and (2) turned in as a written document. Both presentation and written proposal will be graded. In the research proposal, students are required to propose a testable hypothesis developed by using bioinformatics software covered in the class.

- Attendance and participation in class are required.

- Evaluation criteria for this class will be based on completion of the written assignments, and the final presentation.

Course Requirements:

CLRES 2707

Homework assignments	60%
Gene Report	40%

CLRES 2708

Homework assignments	60%
Research Proposal	30%
Research Proposal Presentation	10%

Attendance Policy:

Students are expected to sign-in to each class (computer provided in suite lobby). If a problem is encountered with the sign-in system, please contact the course instructor(s) as well as Lauren Talotta (talottals@upmc.edu) immediately.

Course Grading Scale:

For the computation of the final course grade, the following grading scale will be used:

90-100=A	80-85= B	70-75=C	60-65=D
86-89=B+	76-79=C+	66-69=D+	<60=F

Required Textbook:

There is no required textbook for this class. Reading assignments from selected online articles are recommended to be completed before each class.

Website resources:

Links to lecture-covered Websites on each session will be posted on the Courseweb page.

Academic Integrity:

Students in this course will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity (http://www.provost.pitt.edu/info/ai1.html). Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity. This may include, but is not limited to, the confiscation of the examination of any individual suspected of violating University Policy. Furthermore, no student may bring any unauthorized materials to an exam, including dictionaries and programmable calculators.

Course Schedule

CLRES 2707; Bioinformatics Resources: Data Mining

Date: February 21, 2013

Session 1: Literature Informatics: Beyond PubMed **Chattopadhyay/Iwema**
Next Generation Literature Searching

At the conclusion of this lecture, students will be able to:

1. Formulate complex queries using Medical Subject Heading terms to retrieve relevant literature from PubMed
2. Identify genes associated with a disease
3. Determine mutations that are reported to be linked to a disease or a phenotype
4. Retrieve NIH funded grant information
5. Describe post publication matrix, e.g. Journal impact factors
6. Identify the appropriate journal for manuscript submission
7. Locate appropriate bioinformatics tools for data analysis

Topics:

1. Course Overview
2. Introduction to Medical Subject Headings
3. Overview of next generation literature mining tools
4. Introduction to text similarity searching software
5. Introduction to NIH research funding database
6. Overview of online resources on gene-disease association
7. Overview of online reference management tools
8. Introduction to search engines for life sciences research

Recommended Reading(s):

1. Lu Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. Databas (Oxford) Jan 18

2. Chen YB etal., (2007) The online Bioinformatics Resources Collection. Nucleic Acids Research Jan 35 D780

Homework assignment 1: Answer the assigned questions. Due: March 5

Date: February 26, 2013

Session 2: Genome Biology (part 1) **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Retrieve genome sequence information by formulating complex queries in the organism genome sequence databases
2. Understand, customize & manipulate various display options in the UCSC genome browsers
3. Extract functional information from the annotated genome data
4. Map a region of human genome into other model organisms, such as mouse
5. Place a short nucleotide or protein sequence in the human genome

Topics:

1. Brief overview of genome biology
2. Introduction to organism whole genome sequencing projects
3. Genome sequence databases
4. Genome Browsers : UCSC browser

Recommended Reading(s):

1. Karolchik D, Hinrichs AS, Kent WJ. (2011) The UCSC Genome Browser. Curr Protoc Hum Genet.Oct;Chapter 18:Unit18.6.

Homework assignment 2: Answer the assigned questions. Due: March 19

Date: February 28, 2013

Session 3: Genome Biology (part 2) **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Formulate queries and interpret the result display in MapViewer (NCBI)
2. Understand, customize and manipulate various display options in the Ensembl and GBrowse
3. Retrieve genomic data associated with a genome track in text format, calculate intersections between tracks and fetch DNA sequence covered by a track

Topics:

1. Genome Browsers: Map Viewer , Ensembl, GBrowse
2. Microbial genome: Integrated Microbial Genome
3. Genome Table Browser

Recommended Reading(s):

1. Karolchik D, Hinrichs AS, Kent WJ. (2011) The UCSC Genome Browser. Curr Protoc Hum Genet.Oct;Chapter 18:Unit18.6.

2. Wolfsberg TG. (2011) Using the NCBI Map Viewer to browse genomic sequence data. Curr Protoc Hum Genet. Apr;Chapter 18:Unit18.5.

Homework assignment 3: Answer the assigned questions. Due: March 19

Date: March 5, 2013

Session 4: Gene Information and Regulation **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Retrieve gene related information, such as reference sequences, homologous sequences, gene expression, disease association etc.,
2. Identify promoter sequence of a gene
3. Determine transcription factor binding sites present in a DNA sequence
4. Interpret the Encyclopedia of DNA Element (ENCODE)Project produced genome wide Histone modifications and DNA methylation data and identify promoter, enhancer and silencer sequence(s) present in a genomic region

Topics:

1. Overview of gene-centered information resources
2. Brief overview of gene regulatory elements
3. Promoter databases
4. Transcription factors Databases
5. Encyclopedia of DNA Element Project Data Browser
6. MicroRNA Resources

Recommended Reading(s):

1. Gibney G, Baxevanis AD. (2011) Searching NCBI Databases Using Entrez**.** Curr Protoc Hum Genet. Oct;Chapter 6:Unit6.10.

2. ENCODE Project Consortium, (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011 Apr;9(4):e1001046.

Homework assignment 4: Answer the assigned questions. Due: March 19

Date: March 7, 2013

Session 5: Protein Knowledge Bases (part 1) **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Perform searches in UniProt database and can retrieve a variety of protein related information, such as amino acid sequence, post translational modifications, domain architecture, etc.,
2. Identify interacting partners for a protein of interest

Topics:

1. Overview of protein-centered information gateways
    a. UniProt
    b. Ingenuity IPA
    c. BioBase Protein Knowledge Library
    d. NextBio
2. Protein Domain Databases
    a. InterPro
    b. Molecular Modeling Database
3. Protein Protein interactions
    a. GRID
    b. STRING

Recommended Reading(s):

1. O'Donovan C, Apweiler R. (2011) A guide to UniProt for protein scientists. Methods Mol Biol. 694:25-35.

2. Szklarczyk D, etal., (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011 Jan;39 (Database issue):D561-8.

Homework assignment 5: Answer the assigned questions. Due: March 19

Date: March 19, 2013

Session 6: Protein Structures **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Retrieve secondary structure (helix and turns) information for a protein
2. Retrieve 3D structure of a protein
3. Visualize proteins 3D structure by manipulating display parameters present in a structure viewer software

Topics:

1. Introduction to protein structure databases
    a. Protein Databank
    b. NCBI Structure

2. 3D structure view
    a. Cn3D and FirstGlance

Recommended Reading(s):

1. Gibney G, Baxevanis AD. (2011) Searching NCBI Databases Using Entrez**.** Curr Protoc Hum Genet. Oct;Chapter 6:Unit6.10.

2. Rose PW, etal. (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. Jan;39(Database issue):D392-401.

Homework assignment 6: Answer the assigned questions. Due: March 26

Date: March 21, 2013

Session 7: Genetic Variations and Somatic Mutations in Cancer **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Identify SNPs and CNVs present in a gene sequence
2. Retrieve somatic mutations present in a gene reported in various cancer sub types
3. Retrieve chromosomal aberrations associated with cancers
4. Predict the functional consequences for a mutation

Topics:

1. Brief overview of human genetic variations
2. Introduction to various variation databases
    a. Single Nucleotide Polymorphism – dbSNP
    b. Copy Number Variations – Database of Genomic Variants (DGV) and dbVar
    c. Online Mendelian Inheritance in man (OMIM)
    d. Human Gene Mutation Database
    e. Catalogue of somatic mutations in cancer-COSMIC
3. Functional analysis of mutation
    a. FastSNP

Recommended Reading(s):

1. Forbes SA etal. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. Jan;39(Database issue):D945-50.

2. Ashley EA,(2010) Clinical assessment incorporating a personal genome. Lancet. May 1;375(9725):1525-35.

Homework assignment 7: Answer the assigned questions. Due: March 26

Date: March 26, 2012	
Session 8: Review on Gene Report	**Chattopadhyay/Iwema**

Due March 28: Gene Report

End of CLRES 2707; Bioinformatics Resources: Data Mining

CLRES 2708; Bioinformatics Resources: Data Analysis

Date: March 28, 2013	
Session 1 : Nucleotide Sequence Analysis	**Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Design PCR primers to amplify a DNA sequence
2. Perform *in silico* restriction digestion mapping
3. Create a digital vector map
4. Generate a multiple DNA sequences alignment plot

Topics:

1. Overview of molecular biology Sequence analysis software packages
2. Introduction to CLC Main Workbench
3. PCR primer design
4. Restriction digestion mapping
5. Plasmid map drawing
6. *In silico* cloning

Recommended Reading(s):

1. Untergasser A, etal, (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W71-4. Epub 2007 May 7.

Homework assignment 1: Answer the assigned questions. Due: April 09

Date: April 02, 2013	
Session 2 : Protein Sequence Analysis	**Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Analyze a protein sequence and generate a hydropathicity plot
2. Predict interacting partners for a protein of interest

3. Predict potential kinase specific phosphorylation sites
4. Generate a multiple sequence alignment plot

Topics:

1. Introduction to CLC Main Workbench-protein sequence analysis tools
2. Overview of ExPASy proteomics tools
3. Multiple sequence analysis
4. Protein interaction partner prediction
5. Post translational modification prediction

Recommended Reading(s):

1. Xue Y, etal. (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics. 2008 Sep;7(9):1598-608. Epub 2008 May 6.

Homework assignment 2: Answer the assigned questions. Due: April 09

Date: April 04, 2013

Session 3 : Sequence Similarity Searching **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Understand basic theories behind sequence similarity search algorithms
2. Perform a sequence similarity search against a DNA or Protein databases
3. Interpret and manipulate the BLAST search result by selecting the appropriate algorithms and display parameters available in the NCBI BLAST server

Topics:

1. Introduction to sequence similarity search algorithms
2. Basic Local Alignment Search Tools (BLAST)
    a. PSI BLAST
    b. PHI BLAST
3. Pre-computed BLAST Links database (BLINK)

Recommended Reading(s):

1. Wheeler D, Bhagwat M. (2007) BLAST QuickStart: example-driven web-based BLAST tutorial. Methods Mol Biol. 2007; 395:149-76.

Homework assignment 3: Answer the assigned questions. Due: April 16

Date: April 09, 2013

Session 4 : Introduction to High-Throughput Gene Expression Microarray Data Repositories **Chandran**

At the conclusion of this lecture, the student will be able to:

1. Understand basic concepts present in various gene expression measurement technologies
2. Identify differential expressed genes in a condition of interest
3. Identify conditions where a gene of interest is differentially expressed and download the expression data from microarray database for further statistical analysis

Topics:

1. Introduction to high-throughput gene expression technologies
2. Brief overview of microarray platforms and data types
3. Overview of high-throughput gene expression data repositories
    a. NCBI Gene Expression Omnibus (GEO)
    b. EBI Array Express
    c. Next Bio

Recommended Reading(s):

1. Barrett T, etal. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. Jan;37(Database issue):D885-90.

Homework assignment 4: Answer the assigned questions. Due: April 16

Date: April 11, 2013

Session 5 : Microarray Data Analysis **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Export downloaded gene expression data into a microarray data analysis software package and run powerful statistical tools to analyze and visualize the transcriptomics data and generate a list of differentially expressed genes.

Topics:

1. Overview of microarray data analysis software packages
2. Outline of microarray analysis pipelines
3. Quality control and Data pre-processing
4. Normalization and statistical tests for differential expressions
5. Class discovery, comparison and prediction

Recommended Reading(s):

1. Simon R, etal. (2007) Analysis of gene expression data using BRB-ArrayTools. Cancer Inform. Feb 4; 3:11-7.

Homework assignment 5: Answer the assigned questions. Due: April 23

Date: April 16, 2013

Session 6 : Biological Pathway Analysis (part 1) **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Identify statistically overrepresented biological functions and pathways associated with a list of differentially expressed genes
2. Draw a pathway diagram with gene expression data overlay

Topics:

1. Overview of biological pathway databases
2. Introduction to the Gene Ontology
3. Overview of Gene Set Enrichment Analysis
4. Introduction to pathway analysis software

Recommended Reading(s):

1. Huang da W, (2009)
   Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4(1):44-57.

2. Shmelkov E, Tang Z, Aifantis I, Statnikov A. (2011)
   Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. Biol Direct. Feb 28;6:15.

Homework assignment 6: Answer the assigned questions. Due: April 23

Date: April 18, 2013

Session 7 : Biological Pathway Analysis part 2 **Chattopadhyay**

At the conclusion of this lecture, the student will be able to:

1. Create a protein interaction network map from a list of genes by selecting and manipulating a variety of algorithm parameters available in pathway analysis software

Topics:

1. Introduction to pathway analysis software
2. Protein interaction network map

Homework assignment 7: Answer the assigned questions. Due: April 23

Date: April 23, 2013

Session 8: Review of Students Research Proposals **Chattopadhyay Iwema**