BIOC-6223 Bioinformatics

2 Credits, Tuesday, 4:00 - 6:00 PM
1st class: Monday, August 31
Location: gwu.webex.com/meet/mazumder all Tuesdays.
Office hours: Tuesdays 3-4PM (through appointments)

Course Directors:
Raja Mazumder, PhD, mazumder@gwu.edu
Jonathon Keeney, PhD, keeneyjg@gwu.edu

TA:
Nikhita Gogate: nikigogate@gwmail.gwu.edu

Academic Preparation for Learning: Two – Credit Hour Course: According to the Middle States Commission on Higher Education (MSCHE), students enrolled in a three-credit hour course are expected to engage in a minimum of 75 hours of learning during the period of the course. These hours may include the following: attendance and participation, readings and reflections, working with peers and completing all required assignments for the course.

Academic Integrity: Students are expected to conduct themselves with high ethical standards and comply with the GW’s “Code of Academic Integrity” (see http://www.gwu.edu/~ntegrity/about.html)

Disability: If you have a disability that will require accommodations in this class, register with GW’s Disability Support Services (see https://disabilitysupport.gwu.edu/) and inform the professor. Please allow a reasonable length of time for the necessary accommodations to be arranged.

1. Webex specifics
Here are the call in details. You can call in using a phone or computer.

gwu.webex.com/meet/mazumder
Access code 733 932 866
and use your computer microphone or use "call me" option to connect through phone.

OR

Join by phone if screen sharing is not needed
+1-415-655-0003 US TOLL
733 932 866 Access code

2. All lectures will be videotaped and made available.

3. Attendance during the presentation of the lecture is recommended but not mandatory.

4. Practice problems or examples will be available outside of lecture hours.

5. Students will be able to take exams during their local time zones but with the same time limit restrictions as their counterparts in the EST time zone.

Course Description:
This course will introduce students to application of bioinformatics concepts and methods through the use of molecular biology databases and tools, covering molecular evolution, data flow, functional analysis and analysis communication (e.g. using BioCompute Objects). The course will include lectures,
demonstrations, and practice sessions. Most lectures are followed by hands-on tutorials and some also with take-home assignments to familiarize with the bioinformatics concepts, methods and web resources.

ALL STUDENTS NEED TO BRING THEIR LAPTOPS TO CLASS. IF YOU DO NOT HAVE A LAPTOP LET ME KNOW.

Course Objectives:
By the end of the course, students will be able to
- identify which bioinformatics tools and/or resources to use to perform structure and function analysis of genes/proteins within an evolutionary context;
- use these tools and resources to explore and predict function;
- integrate information from diverse but complimentary resources to answer biological questions.
- communicate analysis workflow instance in a human and machine readable format (BioCompute Objects).

Grading:
Grades for the course will be based on take-home assignments and term project:
Assignments/quiz (20%), midterm (35%), final term project (40%), class participation and attendance (5%).

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100-94</td>
</tr>
<tr>
<td>A-</td>
<td>93-90</td>
</tr>
<tr>
<td>B+</td>
<td>89-86</td>
</tr>
<tr>
<td>B</td>
<td>85-80</td>
</tr>
<tr>
<td>B-</td>
<td>79-75</td>
</tr>
<tr>
<td>C</td>
<td>74-70</td>
</tr>
<tr>
<td>D</td>
<td>69-60</td>
</tr>
<tr>
<td>F</td>
<td>59-below</td>
</tr>
</tbody>
</table>

Expected background: Students should have general knowledge of how to use a computer, common software such as Microsoft Office, and a web browser. One computer science-related course would be useful but not essential.

Use of computers: Please bring your wireless enabled laptops. The GW Blackboard system will be used.

Reading material (optional):
Bioinformatics: A practical guide to the analysis of genes and proteins by Andreas D. Baxevanis and B.F. Francis Ouellette.
# Course Syllabus

(there will be homework/quiz for every lecture)

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topics</th>
<th>Lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Understanding and Communicating Workflows in Bioinformatics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Sep 1</td>
<td>Introduction, databases, PubMed, UniProt, GlyGen</td>
<td>Raja Mazumder</td>
</tr>
<tr>
<td>2</td>
<td>Sep 8</td>
<td>Midterm Challenge outline. JavaScript Object Notation (JSON). Github. BCO Editor.</td>
<td>Jonathon Keeney</td>
</tr>
<tr>
<td>3</td>
<td>Sep 15</td>
<td>Intro BioCompute Objects; BCO creation using BCO Editor. Identification of publications with bioinformatic workflows</td>
<td>Jonathon Keeney</td>
</tr>
<tr>
<td>4</td>
<td>Sep 22</td>
<td>Knowledgebase BCO (OncoMX, BioMuta, BioXpress); BioCompute Objects in Knowledgebases</td>
<td>Hayley Dingerdissen</td>
</tr>
<tr>
<td>5</td>
<td>Sep 29</td>
<td>Machine Learning using MatLab</td>
<td>John/Raja Mazumder</td>
</tr>
<tr>
<td>6</td>
<td>Oct 6</td>
<td>HIVE platform introduction and training/Identification of publications with bioinformatic workflows to make BioCompute Objects.</td>
<td>Jonathon Keeney</td>
</tr>
<tr>
<td>7</td>
<td>Oct 13</td>
<td>Galaxy and BCOs.</td>
<td>Jonathon Keeney and Hadley King</td>
</tr>
<tr>
<td>8</td>
<td>Oct 20</td>
<td>Mid-term Q&amp;A; Including Midterm Challenge draft.</td>
<td>Jonathon Keeney</td>
</tr>
<tr>
<td><strong>Understanding Annotation and Biocuration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Oct 27</td>
<td>Mid-term due. Protein sequence analysis - FASTA, BLAST; Multiple sequence alignment; Gene and protein analysis and knowledge discovery.</td>
<td>Raja Mazumder</td>
</tr>
<tr>
<td>10</td>
<td>Nov 3</td>
<td>Clinical genomics</td>
<td>Harry Burke</td>
</tr>
<tr>
<td>11</td>
<td>Nov 10</td>
<td>Protein sequence, domain and evolution; RACE-P; Family classification and functional analysis. Annotation; Sequence clustering Gene, protein analysis. Molecular phylogenetic analysis</td>
<td>Raja Mazumder</td>
</tr>
<tr>
<td>12</td>
<td>Nov 17</td>
<td>Gene ontology</td>
<td>Darren Natale</td>
</tr>
<tr>
<td>13</td>
<td>Nov 24</td>
<td>G-DOC (cancer genomics)</td>
<td>Yuri Gusev</td>
</tr>
<tr>
<td>14</td>
<td>Dec 1</td>
<td>Q&amp;A. PANTHER, STRING resources</td>
<td>Raja Mazumder</td>
</tr>
<tr>
<td>15</td>
<td>Dec 8</td>
<td>RACE-P functional annotation project due</td>
<td>Raja Mazumder</td>
</tr>
</tbody>
</table>

There will be some mandatory seminar attendance required for this course.

**Bioinformatics platforms (HIVE, Galaxy). Understanding and Communicating Workflows in Bioinformatics**

*Lecture 1*: Introduction; Databases - PubMed, UniProt, GlyGen

How to combine literature search and database search to get a complete understanding of what is currently known about a gene, protein and glycan. Introduction to BioCompute Objects, and capturing a workflow as a BioCompute Object.


Introduction to BioCompute Objects, Github, and BCO creation using BCO editor. JSON: an introduction to JavaScript Object Notation and some hands on exercises formatting data.
Lecture 3: Intro BioCompute Objects; BCO creation using BCO Editor.
How to create BCOs. Identification of publications with bioinformatic workflows. Publication->
BioCompute Object example. An interesting biological question will be posed, along with bioinformatics
analyses that pointed the way, leading to a publication. Students will learn to build a BCO out of the
publication, and will develop an understanding of the context in which the BCO is built and used.

Lecture 4: Knowledgebase BCOs.
Curation in data integration (OncoMX, BioMuta, BioXpress); BioCompute Objects in Knowledgebases
An introduction to the OncoMX knowledgebase, and the BioMuta and BioXpress databases from which it
is built and how to make knowledgebase BCOs.

Lecture 5: Machine Learning
Machine Learning methods using MatLab and metagenomic data.

Lecture 6: HIVE platform introduction and training to make BioCompute Objects.
Hands on with HIVE platform, and identification of publications with bioinformatic workflows to make
BioCompute Objects. An introduction to the High-performance Integrated Virtual Environment (HIVE),
with a hands on tutorial, and the creation of a BCO.

Lecture 7: Galaxy and BCOs.
How to use Galaxy bioinformatics platform to run analysis and make BCOs.

Mid-term Q&A; Jonathon Keeney will be available for the full two hour block to answer questions while
students work on Midterm projects. Midterm due Oct 18th. Submit midterm project after Fall break.

Data Mining and Understanding Annotation and Biocuration

Lecture 9: Protein sequence analysis - FASTA, BLAST; Multiple sequence alignment; Gene and protein
analysis and knowledge discovery
Common sequence analysis pitfalls. How to use information available in gene/protein centric databases
for functional analysis and classification. UniProt, GlyGen, OncoMX resources - How they work and how
they integrate data. How to use the tools for doing genomic and proteomic analysis. How to interpret
flatfiles.

Lecture 10: Clinical genomics
How is bioinformatics used in a clinical setting.

Lecture 11: Protein sequence, domain and evolution; RACE-P; Family classification and functional
analysis; Annotation; Sequence clustering; Gene, protein analysis.
Molecular phylogenetic analysis

Lecture 12: Gene Ontology
The role of Biomedical Ontology in large-scale analysis of data. Why do we need them? How are they
defined? Curation of GO terms.

Lecture 13: G-Doc (cancer genomics platform)
G-DOC. Integration of patient and sample and functional data and tools for knowledge discovery. Cancer
bioinformatics database and tools.

Lecture 14: Q&A. PANTHER, STRING resources for functional analysis.