Introduction to R and Bioconductor

BCHM 695
Syllabus
Summer 2018

Instructor: Dr. Pete E. Pascuzzi
Office: WALC 3053A
Phone: 765-494-3620
email: ppascuzz@purdue.edu (mailto:ppascuzz@purdue.edu)
Hours: TH, 9:30 - 9:50 and 11:40 - 12:00, on Piazza, or by appointment.

Teaching Assistant: Candy Mao
Office: BCHM 211
Phone: 494-0657
email: mao47@purdue.edu (mailto:mao47@purdue.edu)
Hours: TH, 9:30 - 9:50 and 11:40 - 12:00, on Piazza, or by appointment.

COURSE DESCRIPTION

This course provides an introductory, hands-on experience for life science researchers in bioinformatics with R (https://www.r-project.org) and Bioconductor (https://www.bioconductor.org). The various disciplines in the life sciences are generating a wealth of experimental and annotation data. Today's graduate students need experience with modern tools that can help them to access, explore, analyze, interpret and manage the data that they generate in the lab.

Students will use the R programming language and packages from Bioconductor, the R bioinformatics project, as their principal tools for this course. Emphasis will be placed on developing workflows in R that bridge established algorithms for bioinformatics such as limma (https://www.ncbi.nlm.nih.gov/pubmed/?term=25605792), edgeR (https://www.ncbi.nlm.nih.gov/pubmed/?term=22287627) or DESeq2 (https://www.ncbi.nlm.nih.gov/pubmed/?term=25516281). Workflows will included methods to import, QC, transform and visualize genome-scale datasets derived from next generation sequencing experiments.

A critical aspect of bioinformatics that is often inadequate is workflow documentation. This course will use Rmarkdown (https://rmarkdown.rstudio.com), a plain text file format that can integrate computer code, to manage the code and data in complex bioinformatics projects. Rmarkdown files can be exported as HTML, PDF or Word files making it an ideal format for writing reports and documenting analyses. This syllabus was written with Rmarkdown.

The class has lecture and lab components. Lectures will focus on the biological aspects of bioinformatics analysis using recent examples from the literature. In lab, students will work on programming exercises or projects using published datasets. Advanced students will also have the opportunity to work with their own data.

No prior computer programming experience is required, but it is assumed that students can explain the Central Dogma of Molecular Biology (http://www.nature.com/ezproxy.lib.purdue.edu/articles/227561a0.pdf), and have some knowledge of recent advances such as epigenetics and regulatory RNAs.
LEARNING OUTCOMES

- Students will write R scripts that utilize Bioconductor packages for bioinformatic analyses.
- Students will access genome-scale data sets from public repositories and import this data into R for further analysis.
- Students will visualize genome-scale data sets for both quality control and presentation purposes.
- Students will implement strategies to deal with genome-scale datasets including parallel computing.
- Students will be able to critically evaluate the bioinformatic methods and data from publications.
- Students will implement “literate programming” with Rmarkdown to document and share their bioinformatics projects.

TEXTBOOK

No required textbook. However, there are many eBooks on R programming available through the library. Here are a few that are recommended:

R for Data Science (http://r4ds.had.co.nz), Garret Grolemund and Hadley Wickham

LECTURE TIME AND PLACE

<table>
<thead>
<tr>
<th>Session</th>
<th>Day</th>
<th>Time</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>TR</td>
<td>8:40 - 9:30</td>
<td>NLSN 1225</td>
</tr>
<tr>
<td>Lab</td>
<td>TR</td>
<td>9:50 - 11:40</td>
<td>NLSN 1225</td>
</tr>
</tbody>
</table>

ASSESSMENT

This course is offered for a letter grade, although registered audits are welcome. Grades will be determined through successful completion of exercises, projects, a data management notebook, and a final. Final grades will be based on the following:

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eight lab exercises</td>
<td>25%</td>
</tr>
<tr>
<td>Bioinformatic Notebook</td>
<td>25%</td>
</tr>
<tr>
<td>Bioinformatics Project</td>
<td>25%</td>
</tr>
<tr>
<td>Final exam</td>
<td>25%</td>
</tr>
<tr>
<td>Class participation and attendance</td>
<td>0, 2% or 4% bonus points</td>
</tr>
</tbody>
</table>
Exercises will be graded on a scale 0 - 10 with point values clearly indicated. Every effort will be made to give partial credit. Lab exercises and programming projects are not graded on style or efficiency, only on successful completion of the assigned task(s). If your code is well-organized and easy to read, with clear documentation, it is much easier to give you partial credit.

The Bioinformatics Notebook is an open-ended assignment that enables students to synthesize concepts from lecture and lab. The Notebook will be written in Rmarkdown allowing the students to include text narrative, links to web resources, and executeable computer code. The Notebook should include sections ranging from file transfer with Globus, SMB and SCP to quality control metrics for next generation sequencing with clear documentation of how these topics apply to the student’s research. The intent is that students will leave the class with a tangible product that will help them with their research.

Students will work in small teams to develop and complete the Bioinformatics Project. Projects must be approved by the instructor so that they can be completed by the end of the course. Suitable projects will use existing software and data sets. Possible projects are the reanalysis and interpretation of gene expression studies deposited at the National Center for Biotechnology Information or the reuse of ENCODE data.

The final is a traditional exam comprised of short answer questions as well as code chunks that must be completed, edited or corrected.

Perfect attendance is required for 4 bonus points, but one absence is allowed for 2 bonus points. This might seem like a minor reward, but perfect attendance can easily change your letter grade, e.g. 88% (B+) + 4% = 92% (A). Prearranged, excused absences may be allowed at the discretion of the instructor. In either case, active participation in class discussions is expected.

**GRADING SCALE**

<table>
<thead>
<tr>
<th>Score</th>
<th>Grade</th>
<th>Score</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.00 and above</td>
<td>A+</td>
<td>78.00 to 79.99</td>
<td>C+</td>
</tr>
<tr>
<td>92.00 to 97.99</td>
<td>A</td>
<td>72.00 to 77.99</td>
<td>C</td>
</tr>
<tr>
<td>90.00 to 91.99</td>
<td>A-</td>
<td>70.00 to 71.99</td>
<td>C-</td>
</tr>
<tr>
<td>88.00 to 89.99</td>
<td>B+</td>
<td>68.00 to 69.99</td>
<td>D+</td>
</tr>
<tr>
<td>82.00 to 87.99</td>
<td>B</td>
<td>62.00 to 67.99</td>
<td>D</td>
</tr>
<tr>
<td>80.00 to 81.99</td>
<td>B-</td>
<td>60.00 to 61.99</td>
<td>D-</td>
</tr>
<tr>
<td>59.99 and Below</td>
<td></td>
<td></td>
<td>F</td>
</tr>
</tbody>
</table>

**EXTRA CREDIT**

Class attendance is the only way to achieve some extra credit. No exceptions will be made.

**COURSE MANAGEMENT SYSTEM**

We will not use Blackboard for this course. Instead, we will use the Purdue University Research Repository (https://purr.purdue.edu) (PURR). Each student must register for an account with PURR using their Purdue Career credentials, and create a private PURR project with Professor Pascuzzi and the teaching assistant as
collaborators. This project site will be used to turn in all completed assignments and to track your progress in the class. There will be a common PURR project for the course where students will retrieve the lectures, data files and other material as required.

The course also has a Piazza site (https://piazza.com/purdue/summer2018/bchm695/home). Prof. Pascuzzi and Candy Mao will check the site daily, usually between 9:00 and 10:00 AM and again between 4:00 and 5:00 PM. We strongly encourage students to help each other during other times!

**OBTAINING EXTRA HELP**

Professor Pascuzzi will be available to answer your questions immediately after class, during office hours, or by appointment (arranged in class or by e-mail). Alternatively, you can submit questions by e-mail that can be answered in class or by return e-mail. The teaching assistants will be available during office hours, by email or by appointment. If you are struggling, get help!

**ACADEMIC MISCONDUCT**

Academic misconduct of any kind will not be tolerated in any course offered by the Department of Biochemistry. For specifics, please refer to Purdue's Regulations Governing Student Conduct (http://www.purdue.edu/studentregulations/student_conduct/regulations.html)

Any incidence of academic misconduct will be reported to the Office of the Dean of Students. Academic misconduct may result in disciplinary sanctions including expulsion, suspension, probated suspension, disciplinary probation, and/or educational sanctions. In addition, such misconduct will result in punitive grading such as:

- receiving a lower or failing grade on the assignment, or
- assessing a lower or failing grade for the course

Punitive grading decisions will be made after consultation with the Office of the Dean of Students. Please note reported incidences of academic misconduct go on record for reference by other instructors. Further, a record of academic misconduct is likely to influence how current/future situations are handled.

To provide you with an unambiguous definition of academic misconduct, the following text has been excerpted from “Academic Integrity: A Guide for Students”, written by Stephen Akers, Ph.D., Executive Associate Dean of Students (1995, Revised 1999, 2003), and published by the Office of the Dean of Students in cooperation with Purdue Student Government, Schleman Hall of Student Services, Room 207, 475 Stadium Mall Drive West Lafayette, IN 47907-2050.

Purdue prohibits “dishonesty in connection with any University activity. Cheating, plagiarism, or knowingly furnishing false information to the University are examples of dishonesty.” [Part 5, Section III-B-2-a, Student Regulations] Furthermore, the University Senate has stipulated that “the commitment of acts of cheating, lying, and deceit in any of their diverse forms (such as the use of substitutes for taking examinations, the use of illegal cribs, plagiarism, and copying during examinations) is dishonest and must not be tolerated. Moreover, knowingly to aid and abet, directly or indirectly, other parties in committing dishonest acts is in itself dishonest.” [University Senate Document 72-18, December 15, 1972]

More specifically, the following are a few examples of academic dishonesty which have been discovered at Purdue University.

- substituting on an exam for another student
substituting in a course for another student
paying someone else to write a paper and submitting it as one's own work
giving or receiving answers by use of signals during an exam
copying with or without the other person's knowledge during an exam
doing class assignments for someone else
plagiarizing published material, class assignments, or lab reports
turning in a paper that has been purchased from a commercial research firm or obtained from the internet
padding items of a bibliography
obtaining an unauthorized copy of a test in advance of its scheduled administration
using unauthorized notes during an exam
collaborating with other students on assignments when it is not allowed
obtaining a test from the exam site, completing and submitting it later
altering answers on a scored test and submitting it for a regrade
accessing and altering grade records
stealing class assignments from other students and submitting them as one's own
fabricating data
destroying or stealing the work of other students

Plagiarism is a special kind of academic dishonesty in which one person steals another person's ideas or words and falsely presents them as the plagiarist's own product. This is most likely to occur in the following ways:

- using the exact language of someone else without the use of quotation marks and without giving proper credit to the author
- presenting the sequence of ideas or arranging the material of someone else even though such is expressed in one's own words, without giving appropriate acknowledgment
- submitting a document written by someone else but representing it as one's own

CLASS ATTENDANCE

In accordance with University policy, you are expected to attend every scheduled class. If you have a valid reason for missing class such as a University-sponsored activity, religious observances, illness, or family emergency, the instructor or TA will assist you in obtaining information and materials you may have missed. Students who skip class without a valid excuse should not expect the instructor or TA to supply class notes or provide special help. For more information see the Purdue Regulations Governing Classes (http://www.purdue.edu/studentregulations/regulations_procedures/classes.html) and the Class Absence (http://www.purdue.edu/advocacy/students/absences.html) page from the Office of the Dean of Students.

EMERGENCY PREPAREDNESS

In the event of a major campus emergency, course requirements, deadlines and grading percentages are subject to changes that may be necessitated by a revised semester calendar or other circumstances. To get information about changes in this course consult the class PURR site or e-mail or phone the instructor.

ON-LINE COURSE EVALUATIONS
During the last two weeks of the semester, you will be provided an opportunity to evaluate this course and your instructor(s). To this end, Purdue has transitioned to online course evaluations. On Monday of the fifteenth week of classes, you will receive an official email from evaluation administrators with a link to the online evaluation site. You will have two weeks to complete this evaluation. Your participation in this evaluation is an integral part of this course. Your feedback is vital to improving education at Purdue University. I strongly urge you to participate in the evaluation system.

NON-DISCRIMINATION POLICY STATEMENT

Purdue University’s non-discrimination policy will be upheld in this classroom. Purdue University is committed to maintaining a community which recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her own potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life.

Purdue University views, evaluates, and treats all persons in any University related activity or circumstance in which they may be involved, solely as individuals on the basis of their own personal abilities, qualifications, and other relevant characteristics.

For more information, refer to the Purdue Nondiscrimination Policy Statement (http://www.purdue.edu/purdue/ea_eou_statement.html).

CLASS SCHEDULE

Week 01

Lecture 01 Overview of Bioinformatics


Lab 01 Logistics

- Register with PURR and create personal repository
- Check Scholar and Data Depot access
- Overview of RStudio
- Overview of learnR tutorials
- Rmarkdown for reporting
- Getting started on the Bioinformatics Notebook assignment

Lecture 02 Data Management for Bioinformatics

- Discussion of Bioinformatics Notebook
- Discussion of Bioinformatics Project

Lab 02 Managing Projects

• Using Git for version control
• Navigating directories and file paths
• Conversion of data types
• Exercise 01: Data Inventory

Homework

• LearnR tutorials on data structures and data types

Week 2

Lecture 03 Database for Genome-Scale Data

• National Center for Biotechnology Information

Lab 03 Data Manipulation

• Data summaries and transformations
• Data visualization with ggplot2
• Exercise 02: Summarize RNAseq results

Lecture 04 Databases for Genome-Scale Data

• UCSC Genome and Table Browser.
• ENCODE Project

Lab 04 Linking Data

• Data linking and merging
• Data visualization with ggplot 2
• Exercise 03: Data visualization of RNA-seq results

Homework

• LearnR tutorials on data summaries and visualization

Week 3

Lecture 05 Overview of Bioconductor

• F1000 Research Bioconductor Workflows (https://f1000research.com/gateways/bioconductor)

Lab 05 Biological Annotation Data

- Genome annotation packages.
- Biomarts

Lecture 06 Biological Sequence Data


Lab 06 Manipulating Biostrings

- Sequence analysis
- Pattern searching
- Pairwise alignments
- Exercise 04: Sequence extraction and pattern searching

Homework

- LearnR tutorials on Bioconductor
- Initial proposals for bioinformatic project due.

Week 4

Lecture 07 Gene Regulation and Positional Information


Lab 07 In Silico Representation of Genes and Genomes

- Manipulation and annotation of genomic features
- Exercise 05: Identification of Genomic Features

Lecture 08 Epigenetics and Chromatin


Lab 08 Analysis of Genomic Intervals

- Reanalysis of epigenetic data from Julienne, et al.
Homework

- learnR tutorial on Biostrings
- learnR tutorial on GenomicRanges
- Final topic for bioinformatic project approved.

Week 5

Lecture 09 Next Generation Sequencing Projects


Lab 09 Parallel Computing

- Orientation to Purdue Research Computing
- Parallel computing
- Preparing jobs
- **Exercise 06: Unix**

Lecture 10 Limitations of Next Generation Sequencing Projects


Lab 10 Quality Control of NGS Data

- QC of reads
• Mapping reads
• QC of alignments
• Browser tracks from alignments
• Exercise 07: Parallel Computing

Homework
• Tutorial on Unix commands
• learnR tutorial on BAM files

Week 6

Lecture 11 Critical Evaluation of Bioinformatics Results


Lab 11 Differential Gene Expression Analysis with RNA-seq

• RNA-seq analysis with edgeR and DESeq2
• Exercise 08: Reanalysis of Chen, et al data

Lecture 12 Gene List and Pathway Enrichment Analysis


Lab 12 Gene List and Pathway Enrichment Analysis

• Gene Ontology Enrichment Analysis
• Pathway Enrichment Analysis

Homework

• RNA-seq analysis R Notebook.
• Gene list enrichment R Notebook.

Week 7

Lecture 13 DNA-seq Experiments

24877565 (https://www.ncbi.nlm.nih.gov/pubmed/24877565); PubMed Central PMCID: PMC5414828.

- ENCODE Project (https://www.encodeproject.org/)

Lab 13 ChIP-seq Analysis
  - Identifying regions of enrichment from sequencing data
  - Exercise 08: Reanalysis of GEO DNA-seq dataset

Lecture 14 Other NGS Experiments
  - Genome assembly
  - Protein-RNA interactions
  - Chromatin access
  - Metagenomics

Lab 14 Bioinformatics Project
  - Work on bioinformatics project

Week 8
Lecture 15 Open
Lab 15 Open
  - Finish bioinformatic project

Lecture 16 Final
Lab 16 Final