PreMiEr
Precision Microbiome Engineering
NSF Engineering Research Center

This lesson plan was created by a teacher participating in the Research Experiences for Teachers program from the Precision Microbiome Engineering Research Center. Are you interested in spending part of your summer in a lab getting paid to do microbiome research and create lesson plans?
Learn more here: https://premier-microbiome.org/for-teachers-ret/

Lesson plan written by Matthew Hairston

A spotlight on Bioinformatics and Data Science
Teacher Guide

Overview

In the overview:

(1) State the educational goals/learning objectives of the kit.

a. Showcase Bioinformatics/Data Science as a career path.
b. Serve as an introduction to R programming.
c. Demonstrate/Introduce ML and related analysis techniques.
d. Cultivate student interest in interdisciplinary sciences.
e. Improve student engagement through activities that incorporate science relevant to the "real world".
f. Facilitate dialogue/discussion between students in an online classroom.
g. Relate the scope of the science to the students' life experiences.

(2) Summarize the lesson activities.

a. Introduction to Bioinformatics and Data Science
   - Explore the fields of bioinformatics and data science using information collected by students.
     - Have students write down (3) three skills utilized in Bioinformatics and/or Data Science (programming, machine learning, data visualization etc.), (2) two career paths (data engineer, data analyst etc.) and (1) one area of application (drug discovery, personalized medicine, insurance etc.) prior to the lesson.
   - Discuss the interdisciplinary nature of these fields.
     - Biology: Bioinformatics relies on fundamental biological knowledge to understand the underlying mechanisms and functions of biological systems. It involves understanding the structure and function of molecules, genetic inheritance, cellular processes, and evolutionary principles.

    - Computer Science: Computer science provides the computational infrastructure and tools necessary for managing and analyzing large-scale biological datasets. Programming languages, algorithms, database management, and data visualization are key components of bioinformatics.
    - Mathematics and Statistics: Bioinformatics employs mathematical and statistical models to analyze and interpret biological data. Mathematical algorithms help in sequence alignment, phylogenetic analysis, and protein structure prediction. Statistical methods are utilized for data normalization, hypothesis testing, and identifying patterns or correlations within biological datasets.
    - Data Analysis: Bioinformatics utilizes various data analysis techniques to uncover patterns, relationships, and insights from biological data. Machine learning, data mining, and statistical modeling are employed to identify biomarkers, predict protein functions, classify biological samples, and infer biological networks.
    - Genomics and Molecular Biology: Bioinformatics plays a crucial role in analyzing and interpreting genomic data, such as DNA sequencing information, gene expression profiles, and genetic variations. It facilitates the understanding of gene functions, gene regulatory networks, and the impact of genetic variations on health and diseases.

b. Students will have an introductory connective activity within subgroups.
    - First Question:
        - Please tell us your name and three words or phrases that describe your background, and why those words/phrases are important to you.
    - Second Question:
        - What is a memorable experience you have had with a person(s) who is different from you (age, religion, gender, socio-economic, culture, nationality, etc.), and what did you learn about yourself and/or the other person in that experience?

c. Introduction to R Programming:
    - Introduce students to R, a programming language widely used in bioinformatics and data analysis.
    - Familiarize students with RStudio, an integrated development environment for running R code.
        - Note the “Run All Chunks Above” and “Run Current Chunk” buttons in Rstudio. Initiate code up to a point using “Run All..” and after students make modifications to code use “Run Current…”
    - Conduct hands-on activities for students to practice basic R commands for data manipulation and visualization.

d. Exploring Microbiome Data Using R:
    - Guide students through the analysis of a pre-selected microbiome dataset comparing rural and urban samples.
    - Demonstrate how to load the dataset into RStudio and explore its structure.

- Teach students essential R commands to manipulate and visualize the data.
- Encourage students to observe patterns and differences between rural and urban microbiome samples.

e. Applying Machine Learning to Microbiome Data:
   - Introduce students to the concept of machine learning and its application in classifying microbiomes.
   - Discuss the significance of feature selection in achieving accurate classification.
   - Facilitate a discussion on the importance of specific microbial taxa in distinguishing between rural and urban microbiomes.

f. Reflection and Assessment:
   - Engage students in reflective discussions about their learning journey.
   - Provide assessment activities that allow students to demonstrate their understanding, such as creating data visualizations, providing explanations, or writing reports.
   - Ask thought-provoking questions to assess students' comprehension of bioinformatics, R programming, microbiome analysis, and the implications of the research in real-world contexts.

(3) Describe how students will demonstrate their learning.

a. In-class/in-activity discussions – prompts will be provided to guide student-to-student engagement.
   i. For synchronous-online or in-person courses students can communicate in groups.
   ii. For asynchronous-online students communicate using discussion forums.
b. This activity is “low stakes” and is intended to expose students to the science and methodology associated to Bioinformatics. If students struggle, scaffold as necessary.
c. Students will interpret the confusion matrices generated by their code.
d. Students will calculate their model’s accuracy using their confusion matrix.
e. Reflection assignment – students will select up to (3) three questions (differentiate as necessary) from the student guide and address them on an online discussion activity (forum, thread etc.)

(4) List for how many students/groups the materials are intended for:

a. Class up to 28 students
b. 5-7 students per group
c. Up to 4 groups

Learning Outcomes

Students will

1. Gain an awareness of bioinformatics and data science career pathways.
2. Modify code incorporating the basics of R programming.
3. Demonstrate data analysis techniques using R.

4. Discuss the interdisciplinary nature of bioinformatics and real-world applications.
5. Engage in student-to-student collaboration through hands-on activities related to bioinformatics.

Content Standards
This lesson is appropriate for college statistics (MAT-152) students and addresses the following Next Generation Science Standards:

Disciplinary Core Ideas/Practices/Cross-cutting concepts covered
Grades Early-Middle College to Undergraduate

- SLO #1: Organize, display, calculate, and interpret descriptive statistics.
- SLO #2: Apply basic rules of probability.
- SLO #4: Perform regression analysis.

Time Requirements

- Separate prep time: Students ~ 20 minutes | Instructor ~ 60 minutes
- Class-time: 90 minutes
- After class time: Students ~ 15 minutes | Instructor ~ 60 minutes

Materials
Reusable materials:

- Engagement Activity
- Computers or laptops with internet access
- RStudio software installed on each computer
- Handouts with simplified explanations of bioinformatics and data science
- Short, pre-selected biological dataset for analysis (e.g., plant gene expression data)
- Rmarkdown interactive code
- Rmarkdown HTML code
- GitHub repository items

Perishable or disposable materials:
None

Safety
Ensure that students understand and adhere to safe laboratory practices when performing any activity in the classroom or lab. Demonstrate the protocol for correctly using the instruments and materials necessary to complete the activities, and emphasize the importance of proper usage. Use personal protective equipment such as safety glasses or goggles, gloves, and aprons when appropriate. Model proper laboratory safety practices for your students and require them to adhere to all laboratory safety rules.

Background Information

This research paper focuses on the analysis of microbiome data from a study conducted in China, comparing microbiome samples between rural and urban populations. The study aims to explore the potential of machine learning techniques, specifically using 16S sequencing data, to classify microbiomes based on their composition.

Study Design: The researchers collected microbiome samples from individuals residing in both rural and urban areas of China. The samples were processed using 16S sequencing, which provides information about the bacterial taxa present in the microbiome.

Classification of Microbiomes: The study employed machine learning techniques to classify the microbiome samples based on their composition. By training a model on the microbiome data, the researchers aimed to predict whether a sample came from a rural or urban individual.

Importance of Feature Selection: The researchers explored the concept of feature selection, identifying the most important microbial taxa that contributed to the classification accuracy. This allowed them to prioritize specific taxa that have a significant impact on distinguishing between rural and urban microbiomes.

Implications for the Lesson: The research paper serves as the basis for the lesson's dataset, which focuses on analyzing microbiome data from a study comparing rural and urban samples in China. The study demonstrates the potential application of machine learning techniques, specifically using 16S sequencing data, to classify microbiomes based on their composition.

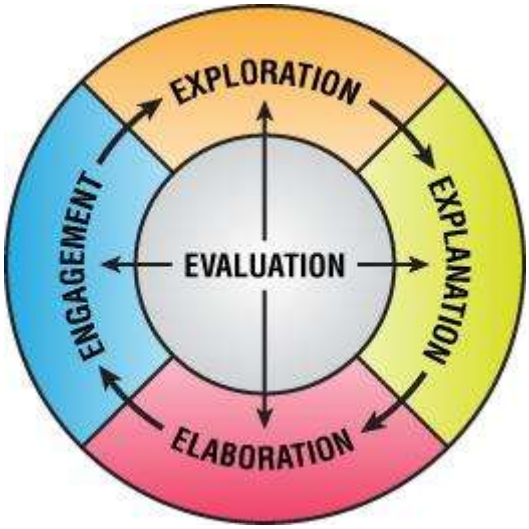
The facilitator should emphasize that while the specific dataset and methods may differ, the paper showcases the broader concept of using machine learning and microbiome data to explore differences between rural and urban microbiomes. It highlights the importance of feature selection and the identification of microbial taxa that contribute to accurate classification.

Preparation
Before Class:

- Prepare the computers or laptops with internet access:
  - Ensure that each computer or laptop is in working order and connected to the internet.
  - Verify that RStudio software is installed on each computer.
  - Test the functionality of RStudio to ensure a smooth learning experience.
- Set up the required materials:
  - Print handouts with simplified explanations of bioinformatics and data science.
  - Prepare a short, pre-selected biological dataset in a suitable format (e.g., CSV or Excel) for analysis.

- Between Classes (While Students are Absent):
- Review the lesson plan:
    - Familiarize yourself with the objectives and flow of the lesson plan.
    - Refresh your understanding of R programming and the activities involved.
- Check the computers and software:
    - Ensure that the computers or laptops are in working order and connected to the internet.
    - Verify that RStudio software is still installed and functional.
- Prepare the dataset and R code:
    - Load the short biological dataset into RStudio and ensure it is in a format suitable for analysis.
    - Create a copy R script or Rmarkdown document(s) and place inside a new folder named “backup”.
    - Test the code snippets to ensure they run smoothly and produce the expected results.
- Print additional copies of handouts (if needed):
    - Check if there are any additional students or copies needed for handouts.
    - Print additional copies of the handouts with simplified explanations of bioinformatics and data science, if necessary.
- Gather any supplementary materials:
    - Collect any additional resources or examples related to bioinformatics and data science.
    - Prepare any extension activities or resources for interested students.

Guiding the lesson using the 5E Learning Cycle



Engage (20 min)

- Start with the engagement activity. Within student groups have students answer the following questions:
    1. Please tell us your name and three words or phrases that describe your background, and why those words/phrases are important to you.

2. What is a memorable experience you have had with a person(s) who is different from you (age, religion, gender, socio-economic, culture, nationality, etc.), and what did you learn about yourself and/or the other person in that experience?

- Begin the lesson by asking students to brainstorm different areas of science where computers and data analysis are used.
- Discuss their responses and introduce bioinformatics and data science as fields that combine biology, computer science, and data analysis.
- Share examples of how bioinformatics and data science contribute to advancements in fields like genomics, drug discovery, and personalized medicine.

Explore (20 min)

- Introduce R programming as a widely used language for data analysis in bioinformatics and data science.
- Briefly explain the importance of RStudio as an integrated development environment for running R code.
- Demonstrate basic R commands for data manipulation and visualization.
- Engage students in a hands-on activity, allowing them to explore RStudio and run simple R commands to manipulate and visualize data.

Explain (25 min)

- Provide students with a short biological dataset, such as plant gene expression data, in a pre-loaded format (e.g., CSV or Excel).
- Guide students through the process of loading the dataset into RStudio and exploring its structure.
- Explain the concept of data analysis and its application in biology, using the dataset as an example.
- Demonstrate simple data analysis tasks, such as calculating summary statistics or creating basic plots, using R code.
- Encourage students to ask questions and discuss their observations and interpretations of the dataset.

Elaborate (15 min)

- Facilitate a class discussion on the real-world applications of bioinformatics and data science in biology. Utilize the following prompts:
    - Can you explain the significance of the ROC curve and the area under the curve (AUC) value in evaluating the performance of the ML model?
    - What do higher AUC values indicate?
    - Can you explain the concept of variable importance in the context of the ML project? How did the variable importance plot provide insights into the iris data set?
    - How can ML techniques and analysis be applied to real-world scenarios beyond the urban vs rural and iris data sets? Can you think of any examples where ML can be used for classification or prediction?

- Encourage students to think critically about how data analysis can provide insights and support decision-making in various biological contexts.
- Provide examples of how bioinformatics can be used to understand diseases, environmental challenges, or genetic variations.
- Engage students in a brainstorming activity where they identify potential research questions or problems that could be explored using data analysis techniques.

Evaluate (5 min)

- Conclude the lesson by summarizing the key points covered, emphasizing the potential of bioinformatics and data science as exciting career paths.
- Encourage students to reflect on their learning and ask any remaining questions they may have.
- Provide students with additional resources, such as websites or online courses, to further explore bioinformatics and data science outside the classroom.

Extend(optional) (5 min)

- Offer extension activities for interested students, such as independent research projects, coding challenges, or exploring online bioinformatics tools and databases.

Differentiated Learning and Extension Activity ideas

Differentiated Instruction:

- Provide optional resources, such as simplified explanations or visual aids, for students who may need extra support in understanding the concepts of bioinformatics and data science. There is an included HTML and PDF file of the activity just in case students are experiencing technological difficulties with the activity.
- Offer additional materials or extension activities for advanced learners to delve deeper into the topics covered in the lesson. Good options for these are modifying plot names/colors, extending the activity by having students calculate other metrics like accuracy, precision, recall etc.
- Provide flexibility in the level of complexity for the hands-on activities, allowing students to choose the level that best suits their abilities.

Collaborative Learning:

- Structure activities that promote collaboration, such as group discussions or peer-to-peer analysis of the microbiome data.
- Encourage students to work in pairs or small groups, where they can learn from and support each other.
- Provide opportunities for students to share their findings, insights, and challenges with their peers, fostering a sense of community and collective learning.

Scaffolding:

- Break down complex tasks, such as loading and manipulating the dataset in RStudio, into step-by-step instructions with clear guidance.
- Provide prompts or templates to assist students in organizing their analysis and interpretation of the microbiome data.

- Gradually release responsibility to students by allowing them to independently explore additional features or conduct further analyses.

Adapted Materials:

- Modify handouts or resources to accommodate diverse learning needs, ensuring clarity of language, font size, and visual elements.
- Incorporate multimedia resources, such as videos or interactive visualizations, to provide alternative ways of understanding the content.
- Consider using a variety of data visualizations to cater to different learning preferences and help students grasp the patterns and insights in the microbiome data.

Flexible Assessment Options:

- Provide various assessment methods, such as written reflections, oral presentations, or data visualizations, allowing students to demonstrate their understanding in different ways.
- Offer choices for the format or topic of final projects or assessments, aligning with students' interests and strengths.
- Focus on the process of learning and growth, providing constructive feedback and opportunities for students to revise and improve their work.

Answers to Questions in the Student Guide

1. What is variable importance?
    a. Variable importance is a way to measure how important each feature or attribute is in a machine learning model. The ranking helps us understand which features have a greater impact on the model's performance and decision-making, regardless of the specific numbers associated with them.
2. Can you explain some basics of R programming? How is R used in data analysis and bioinformatics?
    a. Students can identify syntax, plotting or anything they noticed. If students struggle, guide them by asking syntax-related questions.
    b. R is a programming language commonly used in data analysis and bioinformatics. It helps researchers manipulate data, perform statistical analyses, and create visualizations.
3. What are the key steps involved in analyzing microbiome data using R, from loading the dataset to manipulating and visualizing the data?
    a. The key steps in analyzing microbiome data using R include loading the dataset, manipulating the data, conducting exploratory data analysis, performing statistical analysis, and visualizing the results. Language is great, but the workflow is most important: load > restructure > analyze > visualize.
4. How does machine learning play a role in classifying microbiomes based on their composition, and what are the benefits and challenges of using machine learning techniques in this context?
    a. Machine learning plays a role in classifying microbiomes by learning patterns in their composition. Benefits include accurate classification, but challenges include data preprocessing and model selection.

5. How does the analysis of microbiome data contribute to our understanding of health, environmental factors, and potential connections to diseases or health disparities?
    a. Analysis of microbiome data enhances our understanding of health, environmental factors, and their connections to diseases or health disparities.
6. Reflecting on the hands-on activities and discussions, how has your understanding of bioinformatics, data science, and the analysis of microbiome data evolved? Can you identify any potential real-world applications or future research directions based on what you have learned?
    a. Answers will vary from student to student.

Supplemental Resources

- **Introduction to R Programming**
  https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- **Link to China Urban vs Rural Paper**
  https://tinyurl.com/urban-rural-paper
- **Basic evaluation measures from the confusion matrix**
  https://tinyurl.com/basic-eval
- **ROC curves and AUC Explained (video)**
  https://tinyurl.com/ROC-AUC-explained
- **Visualizations in R**
  https://tinyurl.com/r-visuals-iris

Name _____ **Student Guide**
Date _____

Overview

During this activity, you will explore the exciting fields of bioinformatics and data science. You will learn the basics of R programming and use it to analyze microbiome data from a study comparing rural and urban samples. By the end of this activity, you will gain an understanding of how data analysis techniques can be applied to biology and how machine learning can help classify microbiomes based on their composition.

Background

Bioinformatics and data science combine biology, computer science, and data analysis to study and understand complex biological systems. In this activity, you will focus on the analysis of microbiome data, which involves studying the microbial communities present in different environments or organisms. This knowledge helps us gain insights into various biological processes and their impact on health and the environment.

Materials

- Computers or laptops with internet access
- RStudio software installed on each computer
- Handouts with simplified explanations of bioinformatics and data science
- Short, pre-selected microbiome dataset for analysis

Procedure

Introduction to Bioinformatics and Data Science:

- Learn about the exciting career paths in bioinformatics and data science.
- Understand the interdisciplinary nature of these fields and their real-world applications.

Introduction to R Programming:

- Familiarize yourself with R, a programming language widely used in data analysis and bioinformatics.
- Learn about RStudio, an integrated development environment for running R code.
- Follow along with the facilitator as they demonstrate basic R commands for data manipulation and visualization.
- Engage in hands-on activities to practice using RStudio and running simple R commands.

Exploring Microbiome Data Using R:

- Receive a short microbiome dataset, which includes information about microbial communities.
- Load the dataset into RStudio and explore its structure.
- Learn how to manipulate and visualize the data using R commands.
- Analyze the microbiome data and observe patterns and differences between rural and urban samples.

Applying Machine Learning to Microbiome Data:

- Discover how machine learning techniques can help classify microbiomes based on their composition.
- Understand the concept of feature selection and its importance in accurate classification.
- Learn about the significance of microbial taxa in distinguishing between rural and urban microbiomes.

Questions

1. What is variable importance?
2. Can you explain some basics of R programming? How is R used in data analysis and bioinformatics?
3. What are the key steps involved in analyzing microbiome data using R, from loading the dataset to manipulating and visualizing the data?
4. How does machine learning play a role in classifying microbiomes based on their composition, and what are the benefits and challenges of using machine learning techniques in this context?
5. How does the analysis of microbiome data contribute to our understanding of health, environmental factors, and potential connections to diseases or health disparities?
6. Reflecting on the hands-on activities and discussions, how has your understanding of bioinformatics, data science, and the analysis of microbiome data evolved? Can you identify any potential real-world applications or future research directions based on what you have learned?

p.adjust()
normalization
compositionality