RESEARCH ARTICLE

Design and implementation of semester long project and problem based bioinformatics course [version 1; referees: 1 approved with reservations]

Geetha Saarunya, Bert Ely
Biological Sciences, University of South Carolina, Columbia, South Carolina, 29208, USA

Abstract

Background: Advancements in 'high-throughput technologies' have inundated us with data across disciplines. As a result, there is a bottleneck in addressing the demand for analyzing data and training of 'next generation data scientists'.

Methods: In response to this need, the authors designed a single semester “Bioinformatics” course that introduced a small cohort of students at the University of South Carolina to methods for analyzing data generated through different ‘omic’ platforms using variety of model systems. The course was divided into seven modules with each module ending with a problem.

Results: Towards the end of the course, the students each designed a project that allowed them to pursue their individual interests. These completed projects were presented as talks and posters at ISCB-RSG-SEUSA symposium held at University of South Carolina.

Conclusions: An important outcome of this course design was that the students acquired the basic skills to critically evaluate the reporting and interpretation of data of a problem or a project during the symposium.

Keywords
bioinformatics education, problem-based learning, project-based learning, hands-on course

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the Bioinformatics Education and Training Collection collection.
Corresponding author: Geetha Saarunya (sreeramc@email.sc.edu)

Author roles: Saarunya G: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Ely B: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2018 Saarunya G and Ely B. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero ”No rights reserved” data waiver (CC0 1.0 Public domain dedication).

How to cite this article: Saarunya G and Ely B. Design and implementation of semester long project and problem based bioinformatics course [version 1; referees: 1 approved with reservations] F1000Research 2018, 7(ISCB Comm J):1547 (https://doi.org/10.12688/f1000research.16310.1)

Introduction

Bioinformatics is a rapidly growing interdisciplinary field because of advances in both computer science and the life sciences. Rapid advances in sequencing technologies have led to a deluge of biological data, creating a need for expeditious, efficient, and effective analyses. Practitioners of bioinformatics now add techniques from statistics, information science and engineering to develop algorithms and build predictive models to understand the dynamics within a biological system. This paradigm shift in how bioinformatics is perceived has resulted in an evolutionary model of growth across both of its root disciplines. Bioinformatics as a field also enjoys a degree of duality: “episteme” (scientific knowledge) and “techne” (technical know-how), leading to the idea of “Science informing the tools and the tools enabling science”. In a 2017 survey of 704 NSF principal investigators, more than 90% of respondents replied that they were soon to be working with data sets that required high-performance computing, and they also identified bioinformatics data analyses to be the most urgent and unmet need required for successful completion of their projects. Increased exposure of students at an undergraduate level will help address the need for specialists working in this field and also make the students attractive for opportunities in industry or in graduate school. The Global Organization for Bioinformatics Learning, Education and Training (GOBLET) identified through surveys that the skills required for “basic data stewardship” are taught only in ~ 25% of education programs creating a gulf between theory and practice.

Many courses have been designed and implemented to address the gaps faced in the field. They are project based, problem based or a combination of both to study one or more ‘next-generation’ datasets. The courses have been designed as workshops or as semester long courses using analyses from a single next-generation technology. The authors haven’t come across a course that incorporates multi-omics data analyses in a single semester. There have been studies that address a single problem using multi-omics approaches and there have been pipeline designs that help integrate these data under a single platform.

In response to this need, we designed a single semester course on bioinformatics in the Department of Biological Sciences at University of South Carolina that was targeted towards undergraduate seniors and graduate students who were mainly bench scientists working on experiments which generated data across different ‘omic technologies’ using different living systems.

Challenges in design of bioinformatics curriculum

The curriculum task force of the ‘International Society of Computational Biology’, a scholarly society for both bioinformatics and computational biology research scientists across the world, identified a set of 16 core competencies established through surveys and an iterative process of inputs from people associated with the fields of bioinformatics and computational biology.

However, one of the biggest challenges is the heterogeneity of the backgrounds of the course participants. There is ‘no one size fits all’ while designing a bioinformatics course. In fact, there are three different types of user groups that employ bioinformatics in their research (Table 1), and each of these user groups requires different competencies.

Thus, there was considerable diversity in the backgrounds of the students registered for our course. In response, we chose to follow a ‘learner adaptable’ style of design of the curriculum. This approach allowed us to design the course based on the students’ knowledge of the subject and their expectations of the course.

Methods

Course design

Course conception. This course was designed to provide a structured Bioinformatics course that is geared towards the needs of students working on different “omics” experiments. The general premise of the course was to critically examine and analyze published or in-preparation datasets across different biological systems in a hands-on fashion. In addition, we wanted to introduce the students to the R programming language.

Course Participants. We had nine participants registered for the course. Four of the students were undergraduate seniors, four were first or second year graduate students and one of them was an emergency medical technician (EMT) with a Bachelor of Science degree who was taking additional classes for credit and is now in medical school.

Learning objectives and outcomes of the course. We sent a three-question survey (Table 2) to all the participants to understand their reasoning for registering in the course.

The primary learning objective of the course was to introduce the students to the breadth and depth of the field of Bioinformatics for ‘omics’ data analyses. We also identified the following three course outcomes for the students.

I. At the end of the course, students should be able to identify and implement alternate strategies to answer genomics-based research questions.

II. Students should be comfortable with the use open-source genomic software and command line programming, and be able to use R statistical packages.

III. Students should be able to design and trouble-shoot analyses of nucleotide sequence data and elicit biological information from the data.

Course structure

The course was divided into seven modules spread across the semester: Genome assembly and annotation, Comparative genomics, Introduction to Statistics, Metagenomics, Transcriptomics.
Table 1. Characteristics of user groups.

<table>
<thead>
<tr>
<th>User groups</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics Tool Users (BTU)</td>
<td>These users access bioinformatics resources, packages and software to perform analyses specific to their research domains. e.g. bench scientists, medical professionals</td>
</tr>
<tr>
<td>Bioinformatics Data Scientists (BDS)</td>
<td>These users utilize computational methods to analyze data and advance the scientific understanding of living systems.</td>
</tr>
<tr>
<td>Bioinformatics Engineers (BE)</td>
<td>These users, create, develop and manage novel computational methods needed for novel scientific discoveries.</td>
</tr>
</tbody>
</table>

Table 2. Survey questions sent out to the students.

<table>
<thead>
<tr>
<th>Question premise</th>
<th>Reasons for the question</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1) Previous Programming experience?</td>
<td>We wanted to gauge the level of expertise of the students and identify the level of programming to be introduced in class.</td>
<td>(i) 4 participants had taken a course on R.* (ii) 5 participants had no previous experience using any bioinformatics software or programming languages.</td>
</tr>
<tr>
<td>Q2) Motivation for registering in the course?</td>
<td>We wanted to understand the rationale of the students participating in the course</td>
<td>Unanimous response of the participants was that they were working on some type of benchwork that would generate “omic” data.</td>
</tr>
<tr>
<td>Q3 Take away from the course?</td>
<td>We wanted to ensure our learning outcomes matched the expectations of the course participants.</td>
<td>-Understand types of sequencing technologies -Learn how to analyze data -Learn better practices of biological data management</td>
</tr>
</tbody>
</table>

*Since we did not have this information in the pre-class survey answers, we asked students their experience with programming languages in class. We got 7 responses in total to the pre-lab survey.

Proteomics and Cancer data analysis. Each module ended with a graded research problem either in a prokaryotic system or a eukaryotic system (Table 3 and Supplementary File 1).

Results

Based on the responses of the students, we assigned potential user groups as explained in Table 1 at the start of the class with their expected competency levels at the end of the class. Seven students replied and two students did not reply to the pre-course survey. We were able to obtain permission from six of the seven students who replied to the survey to have their answers published online anonymously. Any identifying information in terms of names or project details have been edited from the responses (Table 4).

Successful completion of the project assigned to every student by the end of a course module determined their competency of the course. In lieu of a final exam, each student designed a research project, conducted appropriate analyses, and summarized their results in the form of a poster or a talk at the end of the semester as part of the ISCB-RSG-SE USA (International society of Computational biology-Regional student group-Southeast USA) conference held on campus on Dec 8/9 of 2017. They also had the opportunity to listen to talks from professors working on bioinformatics projects and interacted with their peers from University of South Florida and University of Alabama. In addition, two graduate students wrote papers on their projects with input from their respective research advisors.

Discussion

This course covered a lot of topics in 13 weeks and some degree of mastery was required for each topic. In addition, half of the students had no familiarity with programming. As a result, many of the students were stretched beyond their comfort zone. However, since this was a small class, we were able to work with the students individually to help them be successful, and also tailor projects to the students’ backgrounds and expectations. An important outcome of this course design was that the students acquired the basic skills to critically evaluate the reporting and interpretation of data of a problem or a project during the symposium.

Our leading goal was to develop a course that was responsive to the needs and background abilities of the participating students. It is important to recognize that every course will have students at different levels of learning with different goals. Hence when designing a course that caters to the needs of the students, it may be a good idea to have a small class.

Dataset 1. Pre-class survey
https://dx.doi.org/10.5256/f1000research.16310.d218863

Dataset 2. Post-class survey
https://dx.doi.org/10.5256/f1000research.16310.d218864
<table>
<thead>
<tr>
<th>Module</th>
<th>Topics covered</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome assembly and annotation</td>
<td>(i) DNA sequencing and its advances over the years. (ii) Assembly of a bacterial genome from nucleotide sequencing data, and submission to NCBI GenBank.</td>
<td>Artemis: A free genome browser and annotation tool that allows visualization of sequence features.</td>
</tr>
<tr>
<td>Comparative Genomics</td>
<td>(i) Strategies to identify prokaryotic and eukaryotic genes. (ii) Strategies for genome comparison: genome size, genomic signature, gene order analyses through sequence alignment.</td>
<td>MAUVE: Multiple genome aligner to compare genomes for evolutionary events and rearrangements.</td>
</tr>
<tr>
<td>Comparative Genomics</td>
<td>Students were introduced to these concepts and then allowed to work on their comparative metagenomics data analyses projects.</td>
<td>R Statistical package: Students were introduced to the R package and were given cheat sheets on how to load, access, and manipulate biological data.</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>(i) Importance of metagenomics across research domains. (ii) Exploring types of research questions answered by metagenomic based studies, data extraction, submission and analyses through MG-RAST pipeline. (iii) Detailed descriptions of the Global Ocean Sampling Expedition data available at the MG-RAST data repository.</td>
<td>STAMP: software package for analyzing taxonomic and metabolic profiles by choosing appropriate statistical techniques.</td>
</tr>
<tr>
<td>Introduction to statistics</td>
<td>(i) Descriptive and Inferential statistics. (ii) Univariate and Bivariate analyses. (iii) ANOVA and PCA.</td>
<td>R Statistical package: Students detected differentially expressed genes using R packages and learned how to be wary of potential sequencing errors.</td>
</tr>
<tr>
<td>Transcripomics</td>
<td>Students were introduced to RNAseq analysis techniques for generating cervical cancer data. They also learned the features of UCSC Cancer genome browser.</td>
<td>UCSC Cancer genomics browser.</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Students were introduced to protein diversity characterization using proteomics. The dataset used for this module was from Bioconductor Conference held at Stanford in July 2016.</td>
<td>R Statistical packages for R/Bioconductor packages to explore, process, visualize, and understand mass spectrometry-based proteomics data.</td>
</tr>
</tbody>
</table>

*All the presentations associated with each module, course assignments and problem assignments are available for access in the supplementary section of the paper. The final projects that were presented as posters and talks are not available for access at this time.*
In our class, every student had a different learning curve. We determined the competency of a student per module by their successful completion of the problem set and or the project. The first objective of the course was to expose the students to not just one living system but many including Bacterial, Human, Drosophila. The other objective was to introduce the students to the R computational platform\textsuperscript{20}. Our initial challenge was to address the problems faced by the students in using the platform for the first time. We wanted the students to understand the intricacies of using R as a programming language but if we repeat this class, we will have the codes for the students as R-markdown documents. We would also have additional R assignments at the beginning of the course and out of class help sessions to help students get comfortable using R.

A major challenge was to identify ways to map the competencies required to the expectations of the course at both the undergraduate and graduate levels. Since we had a small number of students, we designed and delivered a structured curriculum that integrated both the continuously changing and stable technological platforms using model systems that were used by at least one student for every module.

As the important goal of the course was to address the needs of the students, we designed the current model of ‘multi-project’ modules of biological data analyses. Due to the small class size, we were able to give personalized attention to every student. In the future, a big change that we would incorporate would be to separate the projects and problems assigned to graduate and undergraduate students. Generally, the undergraduate students do not have their own data while the graduate students usually have or are in the process of obtaining data that they want to analyze. Therefore, we would either have separate sections for the graduate and undergraduate students or we would have a combined lecture but separate recitation section where the students would apply what they have learned in the lecture portion of the class. The graduate students would be encouraged to develop projects that are relevant to their research while the undergraduates would work in groups on projects designed by the instructor.

### Table 4. Student pre class and expected user groups.

<table>
<thead>
<tr>
<th>Student</th>
<th>Pre-class User group</th>
<th>Expected user group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics Tool User</td>
</tr>
<tr>
<td>2</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics Data scientist</td>
</tr>
<tr>
<td>3</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics tool user, Bioinformatics Data scientist</td>
</tr>
<tr>
<td>4</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics Data scientist</td>
</tr>
<tr>
<td>5</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics Data scientist</td>
</tr>
<tr>
<td>6</td>
<td>Bioinformatics Tool User</td>
<td>Bioinformatics Tool User</td>
</tr>
</tbody>
</table>

### Keypoints
- This course was designed to address the students need to analyze ‘omic’ data sets at University of South Carolina.
- It was divided into seven modules with practical tasks at the end of each module.
- Students designed their projects and presented it as papers, posters and talks at The ISCB- RSG-SEUSA symposium.

### Data availability

Dataset 1: Pre-class surveys 10.5265/f1000research.16310.d218863\textsuperscript{25}

Dataset 2: Post-class surveys 10.5265/f1000research.16310.d218864\textsuperscript{26}

### Ethical considerations

The authors have posted the pre-class survey answers of students who have consented to have their responses published anonymously. All identifying information has been edited from the responses. The post-class survey responses are given as a feedback to the instructors, also anonymously, through an online survey carried out by the university.

### Grant information

The author(s) declared that no grants were involved in supporting this work.

### Acknowledgements

The authors would like to thank Dr. Phillip Buckhaults for the design, conception and delivery of the lectures on “Cancer Genomics”. The authors would also like to thank all the attendees, participants and professors of the Departments of Biological Sciences and Computer Science of University of South Carolina for participating in the first ‘ISCB-RSG-SEUSA’ symposium held this past December of 2017 at Columbia, SC.

### Supplementary material

Supplementary File 1: Course syllabus and teaching materials

Click here to access the data
References


Saarunya and Ely describe a problem-based bioinformatics course designed to meet a need for “next generation data scientists” in the life sciences, a need identified by many current efforts in life sciences education. Case studies of course development efforts like this can be valuable to those seeking to develop similar courses or incorporate those courses into their curricula and looking for ideas or for pitfalls to avoid. The authors do a good service for the field in putting out their efforts and lessons learned in a form from which other educators can benefit. The specific effort here is a nice example of a small project-focused course serving a cohort with some diversity of backgrounds and immediate training needs. While it presents just one small example, that description might reasonably apply to courses many training programs are developing or would like to develop. In addition to the article itself, the supplementary material includes a full syllabus, lecture slides, assignments, and some supplementary materials, increasing its value to others looking to develop course materials in this space.

The authors make a good case for the need for new courses along these lines. They back that need up well with appropriate citations to the relevant literature on life sciences and bioinformatics education. The manuscript provides a good background on prior efforts to characterize the need for bioinformatics training, identify the specific skills required by future life scientists, and how those skills are or are not being provided in practice. The authors further give reasonable consideration to challenges to the design of bioinformatics curricula that they expected to confront in this effort. On the latter point, they might also refer to Williams et al. (2017), which identified a number of other recurring challenges to bioinformatics education in the life sciences. Others in the field might appreciate the perspective of these authors on whether any of the challenges Williams et al. identified were encountered in their effort and, if so, how they were overcome.

The course itself covers a nice range of topics in applied bioinformatics, which might be expected to meet the needs of a diverse set of likely users. The course materials provided in the supplement might therefore find a good audience. One general concern, though, is that the supplementary materials contain some third-party resources, for which it might be more appropriate to include a reference or link rather than the material itself. The teaching approach is fairly applied, with a lot of focus on specific data resources and software, although with some attention to principles behind these resources. While some user communities might favor an approach more grounded in the principles and theory, the focus here seems typical of many bioinformatics courses aimed primarily at biology students. The authors might do a bit more to justify the balance of focus on practice versus theory, with reference to efforts at identifying
specific bioinformatics competencies needed by their likely user community, several of which the paper cites.

The Results present some interesting material in the form of a pre-class survey and post-class course evaluation material. While the cohort here is a single small sample, some useful lessons can be drawn about the diversity of backgrounds and needs of even a small group like this. The paper would be considerably stronger with some more serious assessment of whether the learning objectives of the course were met. That is a non-trivial undertaking and cannot be done retroactively, but might be worth considering for a future iteration of the class if it is being continued. The materials do include results of a university-run course evaluation, which provide some indication of how students felt about the course, although that is different from showing how successfully they learned the material. This post-class evaluation makes for some interesting reading, although if it is being included with the paper, it might bear some comment in the Results and Discussion.

It would be useful also to see some comparison to other similar course material available in publicly accessible forms. While that is a difficult moving target, comparing to a few alternatives from prominent course repositories or MOOCs, particularly to highlight the unusual or especially innovative features of this course, would be valuable.

The paper does a nice job of presenting some lessons learned in the Discussion. It is commendable that the authors spend some time on what did not work so well in this class and consider how it might be done differently in the future. One would ideally like to see this taken further via a more comprehensive formative assessment process – with problems identified via a formal assessment, solutions proposed, and those solutions demonstrated to be effective in a re-assessment. It is understandable that that may be beyond the scope of a one-off paper like this, though, and it is nonetheless easy to see how others developing a class in this domain might benefit from the advice given here to avoid some of the same pitfalls.

Beyond these more specific technical points, the document is clear and generally well-written. I noted just a couple of minor errors:

- p. 4: “International society of Computational biology” should be “International Society for Computational Biology”.
- p. 4: “Regional student group – Southeast USA” should be “Regional Student Group – Southeast USA”.

References

Is the work clearly and accurately presented and does it cite the current literature? Yes

Is the study design appropriate and is the work technically sound? Partly
Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com