frontiers in Bioinformatics

OPINION
published: 23 April 2021
doi: 10.3389/fbinf.2021.656531

Early Requirement for Bioinformatics in Undergraduate Biology Curricula

Matthew G. Niepielko[1] *and Maria Shumskaya*[2]*

[1]New Jersey Center for Science, Technology, and Mathematics, Kean University, Union, NJ, United States, [2]School of Natural Sciences, Biology, Kean University, Union, NJ, United States

Keywords: undergraduate education, computational biology, statistics for biologists, R, stem education

INTRODUCTION

As the world unravels its most impactful event of the century so far – the COVID-19 pandemic, - billions of people turn on televisions, tune into radios, and browse websites trying to understand what the epidemiologic graphs are saying; and in most cases, they turn to media and friends asking to explain what these graphs mean. The COVID-19 pandemic has confirmed: there are huge gaps in the ability for the general population to interpret statistical analyses and graphical representation of biological data (Andrew, 2020; Leybzon, 2020; Tracy, 2020).

In the current situation, understanding data being a prerogative for only data specialists is gone; every health care professional, biologist, chemist, or any natural scientist have taken on the responsibility for the evaluation of massive amount of the pandemic data and delivering the conclusions to their friends and families. But are we really prepared for such work and responsibility? As classically trained in US biologists, we were never required by undergraduate college programs to dive deep into the quantitative analysis world (Cheesman et al., 2007). We were dealing with enzyme kinetics graphs in biochemistry, but not once did we touch "big data"–with exception to some of our friends who were brave enough to enroll into a biostatistics class as an elective. Only later, in graduate school, some had an opportunity to take bioinformatics courses such as computational biology, systems biology, or statistical programming.

Today, there is enough data generated by sequencing, gene expression, bench work on DNA, proteins, and metabolites, thus bioinformaticians have plenty of work to do in phylogenetics, gene expression analysis, genome analysis or an interactome prediction (Hagen, 2000; Gauthier et al., 2019). Many of bioinformaticians either have a computer science background or learned computational analysis in their graduate programs. With the overwhelming amount of data that is available today on any topic, including ecology, biodiversity, epidemiology, we believe that all biologists should receive mandatory training in bioinformatics during their undergraduate years, just as they receive training in organic chemistry or physics. For almost two decades, it has been documented that there is need for undergraduate life science majors to graduate with competency in bioinformatics to help not only scientific progression, but students' careers as well (Bialek and Botstein, 2004; Pevzner and Shamir, 2009; Levine, 2014; American Association for the Advancement of Science, 2015; Sayres et al., 2018), with attempts to implement data science across life sciences curriculum (Dill-McFarland et al., 2021). However, there are still barriers preventing successful inclusion of bioinformatics into undergraduate life sciences education, including lack of student interest, overly full curricula, lack of student preparation, and faculty members belonging to underrepresented groups (Williams et al., 2019). Here, we discuss our opinions and experiences regarding the inclusion of bioinformatics early in undergraduate life science curricula at Kean University, a Hispanic-serving Institution (HSI).

Courses that cover bioinformatics skills should be offered as early as sophomore year, immediately after students complete two semesters of general biology courses and introductory

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Renato Augusto Corrêa Dos Santos, State University of Campinas, Brazil

***Correspondence:**
Maria Shumskaya
mshumska@kean.edu

Specialty section:
This article was submitted to Computational BioImaging, a section of the journal Frontiers in Bioinformatics

Received: *21 January 2021*
Accepted: *08 April 2021*
Published: *23 April 2021*

Citation:
Niepielko MG and Shumskaya M (2021) Early Requirement for Bioinformatics in Undergraduate Biology Curricula. Front. Bioinform. 1:656531. doi: 10.3389/fbinf.2021.656531

statistics. We advocate for this program improvement because biologists need to understand basic data analysis and be familiar with the methods applied, their pros and cons. The lack of important skills necessary to evaluate the validity of a data analysis or draw a critical conclusion from a graph we observe in undergraduate biology majors is devastating. "What is p-value, why do we need it here and what does it tell us? Are graphs scaled correctly for comparison? Do the error bars represent standard error or standard deviation? Is all the data represented on the graph or is there just the averages?"–"How would I know?" These common conversations with students make it clear: the opinion of a person with a college degree in biology can be easily manipulated with some invalid data. We would not expect a biology major to excel in math; rather, we want all students to accrue basic computational analysis skills to deal with the data by embracing the Jim Frost idea: "I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results" (Allison Loves Math Podcast, 2021; Frost, 2021).

Possessing essential computational skills and bioinformatics tools are no longer for special people talented "in computers;" it is a required competency for a biologist (Pevzner and Shamir, 2009; White et al., 2013; Sayres et al., 2018). Computational classes for sophomore level biology majors should focus on biological application, rather than the math theory behind it. A biologist should need to know how and when to use a statistical test and how to interpret the data; the statistical equations behind it should be secondary. Such a course should be taught by a biologist who understands data analysis requirements, who is trained in R, Python, MATLAB, MEGA, PyMOL (Van Rossum and Drake, 1995; MatLab, 2010; PyMOL, 2010; Core R Team, 2017; Kumar et al., 2018) and can design multiple practical exercises for the course. Data on gene expression, heart rates, cholesterol levels, drug efficacy, biodiversity and community structure, evolutionary relationships, selection in a population, and mutations in a gene can be incorporated into class exercises thanks to easily accessible and free online databases and publications.

OUR EXPERIENCE

Computational Courses in a Biology Undergraduate Curriculum

To start filling the gaps in data analysis skills for future biologists, we offer a course on Bioinformatics (3-credits) for sophomores majoring in BS Biology at Kean University. The course is mandatory for the program option in Cell and Molecular Biology and is designed as a set of hands-on exercises on data analysis. Students use Excel and R statistical programming to work with biodiversity data, MEGA to study alignments and phylogenetics, and PyMOL for protein modeling. Introduction to R is taught in an online game form using DataCamp platform (datacamp.com). Since we found that most textbooks are too advanced for our sophomore undergraduates, we developed our own teaching activities. The activities are based on data available from NCBI, NEON, PDB databases, or our own research data (Shumskaya et al., 2019), and some even using the early nucleic acid and protein structure data on SARS-CoV-2 virus (Lorusso and Shumskaya, 2020; Shumskaya and Lorusso, 2020). The activities cited are published online as Open Education Resources to help promote our teaching philosophy surrounding bioinformatics.

For students interested in developing more skills in computational biology, we have developed a minor in Bioinformatics which includes an advanced course on biostatistics and a set of courses on basic computer programming that would count as free electives. This minor works for students majoring in computer sciences or informational technology as well; such students are required to have passed two semesters of general biology and genetics in addition to bioinformatics and statistics.

Additionally, we offer a Bioinformatics and Genomic Science track for our students pursuing a B.S. in Science and Technology in Molecular Biology. Students in this track complete courses in computer programming, statistical programming, and a bioinformatics elective by the end of their sophomore year. The responses from sophomores that complete our undergraduate bioinformatics course and related courses are overwhelmingly positive. In general, none of the students even considered computational biology as a field prior to participating in our bioinformatics course, being unfamiliar with this option. Course surveys reveal that most students become interested in the field when working with data, especially now when a lot of data on current COVID pandemic is available to practice (Johns Hopkins University Center for Systems Science and Engineering (CSSE), 2021) and appreciate learning new software that can help them succeed in multiple courses. A lack of student interest is a clear barrier that prevents students from pursuing higher-level courses that cover topics in bioinformatics (Williams et al., 2019). Our opinion is that a key component in getting students excited and interested in bioinformatics includes hands-on course exercises that cover real, timely, and relevant data (Shumskaya et al., 2019; Lorusso and Shumskaya, 2020; Shumskaya and Lorusso, 2020).

Updating of Pre-Requisite Courses

Biology majors at Kean University are required to take a 3-credits course on statistics. This course traditionally has a lecture component with a heavy math approach; however, at Kean there is an option to add a 1 credit "Probabilistic Methods Lab" taught by a biologist. In this lab, students learn the R statistical programming language in the first half of the semester using the "Undergraduate Guide to R" tutorial (Martin, 2009). Programming activities are designed to reiterate key concepts such as data structures, functions, normalization of data, and graphs using the ggplot R package (Wickham, 2009), followed by the analysis of data such as drug efficacy and RNA expression levels. What separates this lab from a traditional computer programming courses is surrounding algorithms and equations are not detailed; rather, concepts and application of statistical tests using R

are the focus. In this lab, students are "tool users" rather than "tool makers" (Pevsner, 2015). This enables students to analyze data and interpret results without the intimidation of complex equations. Because this course is given during their sophomore year, the exposure to basic computer programming has motivated some students to pursue more advanced computer programming courses during their junior or senior years. This approach helps us introduce bioinformatics into a lower level course, alleviating the "lack of student preparation" barrier (Williams et al., 2019).

Computational Biology for Undergraduate Research

At Kean University, undergraduates can participate in undergraduate research courses [CUREs (Corwin et al., 2015; Rodenbusch et al., 2016; Shortlidge and Brownell, 2016)] such as Research-First-Initiative (RFI, freshmen) or Research Experience in Biology (REB, juniors). Such courses offer students an opportunity to join a faculty-led research project. The 2-3 credits courses are counted as a part of students' 120-college credit program and are often used to jumpstart future independent research projects. Computational biology research projects are offered as part of CUREs. One RFI project focuses on understanding how mRNA localize within a developing cell. Specifically, freshmen are trained in quantifying real mRNA localization data from confocal images using custom MATLAB scripts from (Niepielko et al., 2018), and how to statistically analyze data and compare how mRNA localization changes in various genetic backgrounds. In two back-to-back semesters, students learn biological research such as single molecule *in situ* hybridization and confocal microscopy followed by computational analysis using MATLAB and R statistical programming. One REB course focuses on molecular biodiversity of dead wood decomposing fungi. Students work with Next Generation Sequencing to assess mycobiome gathered from environmental samples, and then employ a bioinformatics pipeline to analyze NGS data. This research course option finishes with students performing ordination and other statistical analyses to study microbial communities identified on dead wood. Research has shown that students engaging in research as undergraduates had the greatest benefit (Russell et al., 2007; Russell et al., 2017). By offering computational biology research and hands on training opportunities built into life science curricula, we believe that this addresses multiple educational barriers including lack of student interest, overly full curricula, and lack of student preparation. Together, we feel that our approach creates an environment that promotes bioinformatics while benefiting students (Levine, 2014) and faculty research projects.

Summer Workshop for High School Students

We believe that bioinformatics and computational biology should be offered to students as early as possible. At Kean University, we offer a bioinformatics workshop for high school students that are interested in any STEM field. The 4 day remote workshop is offered during the summer months as part of Kean University's Group Summer Scholars Research Program. The course is structured with a 2 h morning session and 2 h afternoon session which allows for students to receive a lecture on all the relevant background information in the morning and apply that knowledge by completing hands-on exercises in the afternoon. The hands-on activities cover RNA, DNA, and protein databases, BLAST searches, sequence analysis using MEGA, and protein structure analysis using PyMOL. Based on our experience, introducing bioinformatics to high school students has been overwhelmingly positive. Regardless of their diverse STEM interests, students are receptive to learning about the field and are proficient at completing all the workshop activities which include articulating key findings and developing hypotheses. Although Kean's workshop is not part of a research study, we believe that offering general introduction to bioinformatics at a high school level will help relieve the "lack of student preparation" barrier identified in research studies (Williams et al., 2019). Furthermore, the feasibility and success of the workshop supports our opinion that early exposure to bioinformatics course material should be a strategy integrated into biology curricula and can be accomplished by including

TABLE 1 | Introduction of computational skills in the biology curriculum.

Level	Biology major	Recommended intervention	Course content	Mandatory
Junior high school students	No	16 h summer workshop	Intro into RNA, DNA, and protein databases, BLAST searches, sequence analysis using MEGA, and protein analysis using PyMOL	No
Any college undergraduate level	Yes	Introductory research course on a specific topic, 1–2 credits	Hands-on analysis of real biological data acquired in lab or from publications	No
College freshmen/ sophomores	Yes	1 semester of a lab course on statistics, 1 credit	Basics of R functions, ggplot graphing, interpreting graphs	Highly recommended
College sophomores	Yes	1 semester course on essentials of bioinformatics, 3 credits	Basics of working with biological data in Excel, R, MEGA, PyMOL. t-test, standard deviation and errors, ordination, BLAST, multiple sequence alignment, phylogeny, protein modeling, selection. Online databases NCBI, PDB etc.	Yes
College juniors, seniors	Yes	1 semester course on biostatistics, 3 credits	Advanced data analysis in Excel and R	Yes for minors in bioinformatics; major elective for others

more background information into the course design rather than relying on mandatory pre-requisite courses, which should help alleviate overly full curricula issues.

DISCUSSION

It is no secret that many barriers exist that prevent exposing students to computational biology and bioinformatics, hence introduction of a special course on computational skills into undergraduate biology curricula is in dire need (Sayres et al., 2018; Williams et al., 2019). Our experience shows that the early introduction and a careful planning of computational biology courses has a positive influence on our diverse undergraduate student population. We summarized our steps on introducing computational biology in a biology curriculum in **Table 1**. Our goal is to promote and teach computational skills as early as possible so that students become comfortable with topics such as "How do I analyze data?" "When do I do a certain statistical test?" "What does the p-value actually mean?" In our opinion, biology students learning computer skills from other biologists helps students embrace quantitative biology without fear of overwhelming complex equations and computational algorithms.

From our experience, providing an early opportunity for students to get involved with computational biology spikes their interest to continue to more advanced independent research projects, especially if they participate in CUREs. In a broader sense, such training would have a huge impact on our society. As documented with COVID-19 analyses we discussed above, scientific data can be misrepresented very easily, leading towards rapid spread of misinformation and poor policy choices. The more specialists that receive training in data analysis and data interpretation, the better, regardless of their specialized background. In the future, perhaps general education courses on data analysis and data interpretation can be designed and made a requirement for all student majors.

AUTHOR CONTRIBUTIONS

Both MS and MN contributed equally to writing the manuscript.

REFERENCES

Allison Loves Math Podcast (2021). #36 How to Make Statistics Exciting with Jim Frost. Available at: https://allisonlovesmath.mykajabi.com/blog/JimFrost (Accessed on April 2, 2021).

American Association for the Advancement of Science (2015). Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future. Available at: https://visionandchange.org/about-v-c-chronicling-the-changes/ (Accessed April 2, 2021).

Andrew, G. (2020). Hey, I Think Something's Wrong with This Graph!. Available at: https://statmodeling.stat.columbia.edu/2020/05/18/hey-i-think-somethings-wrong-with-this-graph (Accessed on April 2, 2021).

Bialek, W., and Botstein, D. (2004). Introductory Science and Mathematics Education for 21st-Century Biologists. *Science* 303 (5659), 788–790. doi:10.1126/science.1095480

Cheesman, K., French, D., Cheesman, I., Swails, N., and Thomas, J. (2007). Is There Any Common Curriculum for Undergraduate Biology Majors in the 21st Century?. *Bioscience* 57 (6), 516–522. doi:10.1641/b570609

Core, R. Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Corwin, L. A., Graham, M. J., and Dolan, E. L. (2015). Modeling Course-Based Undergraduate Research Experiences: an Agenda for Future Research and Evaluation. *CBE Life Sci. Educ.* 14 (1), es1. doi:10.1187/cbe.14-10-0167

Dill-McFarland, K. A., König, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y., et al. (2021). An Integrated, Modular Approach to Data Science Education in Microbiology. *PLoS Comput. Biol.* 17 (2), e1008661. doi:10.1371/journal.pcbi.1008661

Frost, J. (2021). Statistics by Jim. Available at: https://statisticsbyjim.com/ (Accessed April 2, 2021).

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A Brief History of Bioinformatics. *Brief. Bioinform.* 20 (6), 1981–1996. doi:10.1093/bib/bby063

Hagen, J. B. (2000). The Origins of Bioinformatics. *Nat. Rev. Genet.* 1 (3), 231–236. doi:10.1038/35042090

Johns Hopkins University Center for Systems Science and Engineering (CSSE) (2021). COVID-19 Data Repository. Available at: https://github.com/CSSEGISandData/COVID-19 (Accessed March 1, 2021).

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35 (6), 1547–1549. doi:10.1093/molbev/msy096

Levine, A. G. (2014). An Explosion of Bioinformatics Careers. *Science*. 344 (6189), 1303–1306. doi:10.1126/science.344.6189.1303

Leybzon, D. D. (2020). Bad Data Visualization in the Time of COVID-19. Available at: https://medium.com/nightingale/bad-data-visualization-in-the-time-of-covid-19-5a9f8198ce3e (Accessed on April 2, 2021).

Lorusso, N. S., and Shumskaya, M. (2020). Online Laboratory Exercise on Computational Biology: Phylogenetic Analyses and Protein Modeling Based on SARS-CoV -2 Data during COVID -19 Remote Instruction. *Biochem. Mol. Biol. Educ.* 48 (5), 526–527. doi:10.1002/bmb.21438

Martin, T. (2009). *The Undergraduate Guide to R*. Princeton, NJ: Princeton University.

MatLab (2010). *Natick*. Massachusetts: The MathWorks Inc.

Niepielko, M. G., Eagle, W. V. I., and Gavis, E. R. (2018). Stochastic Seeding Coupled with mRNA Self-Recruitment Generates Heterogeneous Drosophila Germ Granules. *Curr. Biol.*, 28(12), 1872–1881. doi:10.1016/j.cub.2018.04.037

Pevsner, J. (2015). *Bioinformatics and Functional Genomics*. Hoboken, United States: Wiley.

Pevzner, P., and Shamir, R. (2009). Computing Has Changed Biology-Biology Education Must Catch up. *Science*. 325 (5940), 541–542. doi:10.1126/science.1173876

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., and Dolan, E. L. (2016). Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees. *CBE Life Sci. Educ.* 15 (2), ar20. doi:10.1187/cbe.16-03-0117

Russell, J. E., D'Costa, A. R., Runck, C., Barnes, D. W., Barrera, A. L., Hurst-Kennedy, J., et al. (2017). Correction for Bridging the Undergraduate Curriculum Using an Integrated Course-Embedded Undergraduate Research Experience (ICURE). *CBE Life Sci. Educ.* 16 (1), co3. doi:10.1187/cbe.14-09-0151-corr

Russell, S. H., Hancock, M. P., and McCullough, J. (2007). THE PIPELINE: Benefits of Undergraduate Research Experiences. *Science* 316 (5824), 548–549. doi:10.1126/science.1140384

Sayres, M. A. W., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics Core Competencies for Undergraduate Life Sciences Education. *PLOS One.* 13 (6), e0196878. doi:10.1371/journal.pone.0196878

Shortlidge, E. E., and Brownell, S. E. (2016). How to Assess Your CURE: A Practical Guide for Instructors of Course-Based Undergraduate Research Experiences †. *J. Microbiol. Biol. Educ.* 17 (3), 399–408. doi:10.1128/jmbe.v17i3.1103

Shumskaya, M., and Lorusso, N. (2020). Introduction to Nucleotide Sequence Analysis and Protein Modeling in MEGA and PyMol Using Coronavirus SARS-CoV-2. *QUBES Educ. Resour.* doi:10.25334/37N4-SW29

Shumskaya., M., and Zambell, C. (2019). NMDS to Study Dead Wood Fungi Communities in Parks of New Jersey. NEON Faculty Mentoring Network. *QUBES Educ. Resour.* doi:10.25334/A2ME-QH70

The PyMOL Molecular Graphics System (Version 1.3r1 edu). (2010). *Schrödinger, LLC*, https://pymol.org/2/.

Tracy, S. (2020). COVID-19 in Charts: Examples of Good & Bad Data Visualization. https://analythical.com/blog/covid19-in-charts (Accessed on April 2, 2021).

Van Rossum, G., and Drake, J. F. L. (1995). *Python Reference Manual*. Amsterdam: Centrum voor Wiskunde en Informatica.

White, H. B., Benore, M. A., Sumter, T. F., Caldwell, B. D., and Bell, E. (2013). What Skills Should Students of Undergraduate Biochemistry and Molecular Biology Programs Have upon Graduation?. *Biochem. Mol. Biol. Educ.* 41 (5), 297–301. doi:10.1002/bmb.20729

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Williams, J. J., Drew, J. C., Galindo-Gonzalez, S., Robic, S., Dinsdale, E., Morgan, W. R., et al. (2019). Barriers to Integration of Bioinformatics into Undergraduate Life Sciences Education: A National Study of US Life Sciences Faculty Uncover Significant Barriers to Integrating Bioinformatics into Undergraduate Instruction. *PLoS One*. 14 (11), e0224288. doi:10.1371/journal.pone.0224288

Conflict of Interest: The authors declare that the work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.