Establishing Bioinformatics at SIUC

Principal Investigator: [redacted]

Abstract

The research objective of this proposal is to promote student success in the discipline of bioinformatics through scientific innovation and leadership at SIUC. Bioinformatics is a discipline that uses computational programs and hardware to analyze vast collections of DNA/RNA from next generation sequencing (NGS) for the purpose of surveying or discovery of genetic architecture and function within biological organisms. The field of bioinformatics is quickly growing to be the most important research specializations, but the lack of academic opportunities makes it difficult to prepare students for the job market. Southern Illinois University Carbondale (SIUC) is developing Bachelor's and Master's Degree Programs in Bioinformatics utilizing courses from the departments of computer science, plant biology, zoology, and microbiology. While courses outline the application of NGS technology are taught, methodologies for processing the data are not. When commercial services are used to process sequence data, accuracy and independence are sacrificed. *This proposal seeks to grow and sustain robust expertise in bioinformatics at SIUC.* To achieve this, a team of students and professors will attend a three-day workshop in Detroit, Michigan, that teaches use of Mothur, the major programing software package used to process NGS sequence data. Bioinformatic training will enrich the ability of both students and professors to advance their research by obtaining deeper insights from sequencing technologies to solve complex problems. This will result in publications, new grant submissions, and improve the quality of education for students at all levels benefiting new disciplines in development that depend on expertise in bioinformatics.

Establishing Bioinformatics at SIUC

Principal Investigator: [redacted]

1. Project Objective

Bioinformatics is a discipline that uses computational programs and hardware to analyze vast collections of DNA/RNA from next generation sequencing (NGS) for the purpose of surveying, prediction, or discovery of genetic architecture and function of biological organisms. The field of bioinformatics is quickly growing to be one of the most important research specializations, but the lack of academic opportunities makes it difficult for students to prepare for the job market. Only twenty-two universities in the United States offer bioinformatics degree programs (Universities.com), which presents a tremendous opportunity for student recruitment. Bioinformatics professionals are in high demand within academia, research, and industry. Disciplines such as agricultural chemistry, pharmaceuticals, ecology, and medical research benefit from the application of bioinformatics. Transparency Market Research predicts that the bioinformatics market will reach $30.87 billion by the end of 2020 [1]. *The research objective of this proposal is to promote student success in the discipline of bioinformatics through superlative student training with research innovation and leadership at Southern Illinois University of Carbondale (SIUC).*

This goal has two aspects. The first steps toward developing Bachelor's and Master's Degree Programs in Bioinformatics have been initiated at SIUC (RME filed under School of Computing at SIUC). A student striving to acquire a degree in the Bioinformatics Program will need to take classes in biology, computer science, chemistry, and mathematics. While courses that teach applications of bioinformatics exist at SIUC, a second critical need is missing: No fundamental instruction exists using either of the two major programing software packages for bioinformatic analyses. The objective of this proposal is for training professors and students to use the programing script language, Mothur. It is worth noting that Mothur is currently loaded onto The BigDawg High Performance Computing Cluster (HPCC) at SIUC, which is free of charge to students and professors. *This will establish a growing and sustainable expertise of microbial and ecological bioinformatics at SIUC. Mothur training will increase the number of skilled instructors and students to disseminate bioinformatic skills to future generations of SIUC students.* The following aims target specific goals to achieve this proposal's directive:

Specific Aim #1 **Researchers from SIUC (professors and students) will attend a bioinformatics workshop to acquire formal training using the script line programing language Mothur.**

Specific Aim #2 **Select microbial environments for next generation sequencing to be processed by Mothur, providing data for future research publications and classroom exercises.**

The intent is to use a highly collaborative group (Departments of Microbiology, Plant Biology, Zoology, and Geology) involved in environmental microbiology research projects to enhance SIUC's ability to meet the challenge of providing quality education in bioinformatics. Official training in Mothur will enrich the ability of both students and professors to advance their research by obtaining deeper insights from cutting-edge sequencing technologies to solve complex problems.

2. Background

[redacted]

While the ability to sequence DNA and RNA is a powerful tool, improvements of design and methods continuously increase the volume (by a factor of 10,000) of generated data, forcing new refinements of processing and interpretation [4, 5]. The rapid development of sequencing technology has created a demand for researchers who possess combined knowledge of life science theory, computational software, and biotechnology. A bioinformaticist develops a 'pipeline' of computer language, coded so that it produces an interpretable output. Bioinformatics training is expected to reach across genomics, proteomics, metabolomics, and transcriptomics of all biological life. The greatest professional hurdle is the lack of degree programs and training opportunities at universities [6]. While a rigorous background in computer science is helpful, this is not an absolute requirement; it is more important to understand the 'central hypothesis' (DNA → RNA → Protein) that governs all living things on this planet and upon which all bioinformatic strategies rely. SIUC's current bioinformatics education is limited to the departments of Plant Biology and Zoology, while in Microbiology it is nearly non-existent. Only the application of NGS technology and its benefits to research are taught. Methodologies for processing raw data, including software packages and sequencing platforms, are not taught in any of the departments. Moreover, the professors do not feel confident teaching fundamental bioinformatic development or would rather wait until software producers generate a graphical user interface (GUI). Faced with rapidly changing sequencing technology and emerging applications, it is vital to prepare SIUC for bioinformatics degrees so that students and professors can make novel discoveries now, becoming future leaders, and not future followers.

I used startup funds for a strategic purchase in the [redacted]; an AMD 16 core 'Thread Ripper' CPU computer with solid-state technology for processing large datasets. While this hardware can support processing large data sets, which is valued currently at tens of millions of sequenced DNA reads across dozens of sampled sites. The current level of expertise in my laboratory without Mothur training supports processing only small data sets, that of only hundred thousand DNA sequences across no more than five samples sites. Nonetheless, I have an excellent example of how bioinformatics immediately demonstrated clear positive outcomes for student success at SIUC. Our capabilities were tested on an outreach project involving Carbondale High School students, who were recruited during my Science Café presentation in Spring 2017. The high school students designed and implemented an experiment with my assistance (**Figure 1**). Their research focused on the interaction of subsurface microbes with solid surfaces, and how that interaction affects microbial diversity. DNA was extracted for NGS sequencing (Argonne National Laboratory), the resulting dataset was processed using a rudimentary Mothur programed script designed by a graduate student (Mr. Rui Xiao, 2018

graduate, is a research assistant at U. of Florida School of Medicine). This bioinformatics study consisted of 226,000 DNA reads (relatively small), yet the results suggested a solid research strategy to one of my undergraduate microbiology majors [redacted]. Who recently won the Research Enriched Academic Challenge (REACH) award for her research on a novel phylum of bacteria 'Atribacteria', This small project is only one example of how the ability to process bioinformatic data can lead to career advantages for students of all academic levels at SIUC. [redacted] research will generate an important species description manuscript submission for peer-reviewed publication.

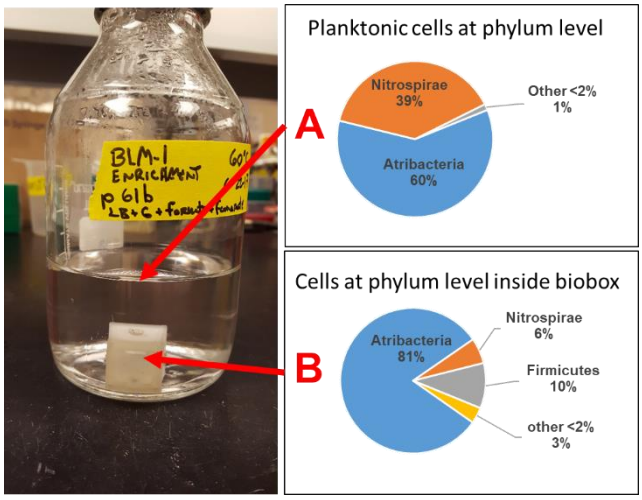

Figure 1. Carbondale high school students' experiment uses Morris Library 3D printed 'bio boxes' to test how solid surfaces influence subsurface microbial diversity. A) Total DNA extracted from planktonic cells resulted in a predominantly binary culture, while B) the bio box exhibited more diversity within the artificial printed structure.

The objective of this proposal is to enhance the experience and research ability of professors and students to teach fundamental bioinformatics, specifically in Microbiology, but also to assist the departments of Zoology and Plant Biology. Several cross-departmental biological research laboratories are perfectly positioned to elevate SIUC's bioinformatics expertise using Mothur as the script language to process NGS datasets. The following SIUC research groups will assist with this proposal's specific aims: the Hamilton-Brehm laboratory (MICR) focuses on biofuels/remediation/ancient DNA and subsurface environments; the Brooks laboratory (ZOOL) focuses on the effect of biogeochemical changes on bacterial communities of freshwater, acid-mine drainages, and geothermal ecosystems; the Jiménez laboratory (ZOOL) focuses on the ecological and evolutionary causes that determine the distribution of bacterial pathogens in parasites; the Rader laboratory (MICR) focuses on beneficial microbiome interactions involved in symbiosis or head trauma. The impetus of inclusion and cohesion of these departments come from our expected merger under the same school. With the promise of new degreed programs in bioinformatics and other programs that are expected to draw from bioinformatics (e.g. Forensic Science), SIUC's investment in this proposal will ensure that students obtain a well-rounded quality education from these exciting new programs to come.

3. Research Plan

These two aims outline the planning and the expected progress needed to meet the overall purpose of this proposal: *This proposal seeks to establish a growing and sustainable expertise and to complement the need of microbial bioinformatics at SIUC that can also inform deeper understanding in multidisciplinary ecological studies.*

3.1. <u>Specific Aim #1</u>: Researchers from SIUC (professors and students) will attend a bioinformatics workshop to acquire formal training using the script line programing language Mothur.

Many different bioinformatics software packages are available (another factor compounding the difficulty of learning bioinformatics) that can be used to process NGS. The two software packages that are most common in academia, open source and made available free, are Qiime2 and Mothur [7, 8]. Dr. Patrick Schloss and his software development team in the Department of Microbiology & Immunology at The University of Michigan initiated the script line programming language called Mothur. They seek to develop a single piece of open source, expandable software to fill the bioinformatics needs of the microbial ecology community. Mothur is currently the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. Performance of Mothur vs Qiime2 is comparably the same. The choice of software depends primarily on which language was introduced first to the user. Because of my experience and success with previous environmental studies using Mothur, this proposal focuses on training professors and students to use the Mothur script language.

3.1.1. Three day workshop training on Mothur in Detroit, Michigan.

In order to increase our growing proficiency using Mothur, formal training must reinforce and expand upon what was previously learned through self-teaching. Dr. Schloss and his staff offer yearly workshops that teach Mothur programing for novice or experienced users. Their three-day workshop addresses both basic and complex problems (see budget justification for three-day itinerary). During the three days, one-on-one sessions with Dr. Schloss and his staff are also available for specific project discussions. The December 17-19 Mothur workshop fits perfectly with student and professor schedules since the fall semester of SIUC will be completed. Four people supported by this proposal will attend the meeting. Dr. Scott Hamilton-Brehm and Dr. Marjorie Brooks will represent faculty interests, and two students whose projects depend upon bioinformatics training will be preferentially considered. Currently, the Hamilton-Brehm and Brooks laboratories are collaborating on three projects investigating how environmental biogeochemistry can alter the structure and function of microbial communities in acid mine drainage landscapes, urban lakes receiving excessive nutrient runoff, and high-temperature geysers. Dr. Agustin Jiménez plans to attend using his own funds. Other students supported from an ongoing NSF STEM scholarship program for biodiversity in the Department of Plant Biology will also participate, supported by their own funding. Those invited graduate and undergraduate students must have a current project using NGS. Dr. Schloss encourages participants of the training session to work on their own data as they learn. Including professors and graduate/undergraduate students from different departments facilitates professional collaboration that will ensure the growth and success of bioinformatics at SIUC.

3.1.2. Post-workshop reinforcement of bioinformatics learning.

To facilitate and strengthen the knowledge acquired from the workshop, all participants will process their own NGS datasets using Mothur to advance their own individual projects. The Hamilton-Brehm research team will focus on datasets from three sources: 1) a doctoral student biofuel project to determine microbial groups that convert lignin to aircraft fuel; 2) samples taken from a geothermal site located in Northern Nevada, in order to generate preliminary results for a NASA Astrobiology proposal due in April, 2019; 3) In collaboration with the Brooks laboratory, an acid mine experiment on the diversity and phenotypic plasticity of soil microbes to cycle

nutrients. Also, the Brooks research team is investigating: 1) How nutrients as selective agents determine the diversity and state of toxic algae present in SIUC's Campus Lake for optimal remediation methods; and 2) Shifts in microbial diversity and ecosystem productivity in streams in Yellowstone National Park depending on beaver (*Castor canadensis*) activity. A joint project between the Hamilton-Brehm and Rader laboratories will investigate microbial populations from Hawaiian Bobtail Squid (*Euprymna scolopes*) cohorts in search of a suspected pathogenic bacterium that causes squid decline. The Jiménez laboratory is working on an inventory of ticks, screening for bacterial pathogens they carry that cause disease in humans and animals.

These projects will help cement skills learned from the workshop and accelerate the research of undergraduate and graduate students so that they can present their data at the SIUC Annual Graduate Student Research Symposium, publish their findings in peer-reviewed journals, and graduate with research experience in bioinformatics. The NGS projects will be monitored (through their PIs) for completion of their research, resulting in a thesis or dissertation, and peer reviewed publication. Undergraduates in this program will be encouraged to continue their bioinformatics training in graduate programs at SIUC. The PIs of this proposal will actively provide training, support, and opportunities to students, demonstrating SIUC's commitment to higher education, research, and career preparation. Mothur is currently loaded onto The BigDawg High Performance Computing Cluster (HPCC) at SIUC, which is free of charge to students and professors

3.2. Specific Aim #2: Select microbial environments for next generation sequencing to be processed by Mothur, providing data for future research publications and classroom exercises.

Bioinformatics is an iterative process of learning and testing on many NGS datasets to gain proficiency. It is also important to be familiar with the origin of DNA, so as to be able to understand when inaccurate results occur from the data. While example datasets and scripts are available, there is no substitution for 'real' data that will always contain unique unplanned problems needing to be resolved. All participants in this proposal will work on their own project data, fundamentally testing each researcher beyond textbook learning to address unforeseen hurtles. We have budgeted for an NGS ninety-six well plate from Argonne National Laboratory using an Illumina MiSeq platform. This type of NGS will allow up to ninety six samples of submitted DNA to be amplified by the bacterial 16S rRNA gene that is typically used in taxonomic identification. Each sample could generate up to 400,000 DNA sequences, multiplied by ninety-six submitted samples, and could result in 38 million DNA reads to be processed by Mothur. This will provide a dataset large enough to meet all research to teaching needs.

3.2.1. Sequencing microbial environments and data processing for research, classroom teaching, and publications.

A few microbial environments that have been sequenced in the Hamilton-Brehm laboratory will provide the initial learning datasets for use in the workshop. Using skills acquired in the workshop, these datasets will be developed into research manuscripts for peer-review. One example is dataset that is through a collaboration with Dr. Brooks (Department Zoology), funded by the Illinois Department of Natural Resources (DNR). Undergraduate Trevor Murphy is measuring the impact of contamination upon ecosystems by using microbial diversity as a diagnostic measure. The research site for this project is a local remediation bioreactor at an acid

mine. If this project is successful, this methodology could be used by the DNR to assess ecological damage via measuring microbial diversity in other sites. Dr. Brooks is using bioinformatics to process NGS datasets collected from Campus Lake to survey the cyanobacterial populations as a means to monitor the efficiency of remediation strategies. Another NGS project comes from Dr. Jiménez, who is working on a screening method for bacterial pathogens carried by ticks. Only a bioinformatic screen will help, as each tick serves as a host to thousands of bacteria and viruses. This project directly impacts public health as cases of Lyme disease, Rocky Mountain Spotted Fever, and Lone Star tick induced meat allergies are on the rise. The requested funds for an NGS sequencing plate from Argonne National Laboratories will successfully launch these projects and a number of others.

The data acquired from ongoing and future NGS projects will be used as a teaching tool in the SIUC's course Geomicrobiology. This course focuses on how microorganisms interact with geochemical processes in nature. As an example of how microbes can be utilized to determine geochemical properties of an environment, each student will be given an NGS dataset and instructions on how to use Mothur to process the data. This will be the first course at SIUC to provide students with a practical 'hands-on' learning experience in bioinformatic development.

4. Timeline

Table 1. Timeline for proposal activities

Tasks	Dec	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct
3.1. Specific Aim #1: Researchers from SIUC (professors and students) will attend a bioinformatics workshop to acquire formal training using the script line programing language Mothur											
3.1.1. Travel to Detroit Workshop	x										
3.1.2. Post workshop reinforcement exercise		x	x	x	x						
3.2. Specific Aim #2: Select microbial environments chosen for next generation sequencing to be processed by Mothur will provide data for future research publications and classroom exercises.											
3.2.1. Sequencing microbial environments and data processing for research, classroom teaching, and publications		x	x	x	x	x	x	x	x	x	x

5. Overall Expected Outcomes

All datasets and projects outlined in this proposal are anticipated to be published in peer-reviewed journals (~6 manuscripts). This will set into motion a growing and sustainable expertise in the field of bioinformatics in the departments of Zoology and Microbiology at SIUC. As the bioinformatics degree offered at SIUC becomes a reality, it is hoped that an RME can be written to create a new course under Microbiology that focuses on bioinformatic software and hardware design with NGS dataset processing. SIUC's investment in bioinformatics will improve the quality of education for students at all levels and will benefit new disciplines in development. The preliminary datasets acquired from the Argonne National Laboratory NGS sequencing plate and Mothur training will greatly enhance all participating professors' external grant submissions chances of success.

References

1. Bioinformatics Market to reach US$30.87 Billion by the end of 2020; Transparency Market Research Albany, NY: GLOBE NEWSWIRE; 2016. Available from: https://globenewswire.com/news-release/2016/09/28/875370/0/en/Bioinformatics-Market-to-reach-US-30-87-Billion-by-the-end-of-2020-Transparency-Market-Research.html.
2. Moser DP, Hamilton-Brehm SD, Fisher JC, Bruckner JC, Kruger B, Sackett J, et al. Radiochemically-Supported Microbial Communities: A Potential Mechanism for Biocolloid Production of Importance to Actinide Transport. Desert Research Institute, Nevada University, Reno, NV (United States), 2014.
3. Hamilton-Brehm SD, Hristova LT, Edwards SR, Wedding JR, Snow M, Kruger BR, et al. Ancient human mitochondrial DNA and radiocarbon analysis of archived quids from the Mule Spring Rockshelter, Nevada, USA. PloS one. 2018;13(3):e0194223.
4. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135.
5. Hutchison III CA. DNA sequencing: bench to bedside and beyond. Nucleic acids research. 2007;35(18):6227-37.
6. Craig DW. Opinion: We Must Make Data More Accessible for Bioinformatics Training 2018. Available from: https://www.the-scientist.com/thought-experiment/opinion-we-must-make-data-more-accessible-for-bioinformatics-training-29861.
7. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology. 2009;75(23):7537-41.
8. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7(5):335.

Establishing Bioinformatics at SIUC

Principal Investigator: [redacted]

Detailed Budget

Line Item	Cost
Mothur three-day workshop training Dec 17 to 19, 2018	
Hotel (La Quinta Hotel, Detroit) $100 per night for three nights	$300/person
Registration (includes breakfast and lunch)	$500/person
Plane Flight (St. Louis to Detroit)	$400/person
Estimated per diem food (dinner and taxis)	$100/person
Total per person	$1300/person
Two professors and two students	4 people
Subtotal for training costs for 4 people	$5200
Argonne National Laboratory: Next generation sequencing plates of 96 samples each: 2x151bp MiSeq run with barcoding and PCR	$3500
DNA extraction, reagents, sample collection supplies	$1300
Total budget	$10,000