Personal, Relevant Background, and Future Goals Statement for NSF GRFP

Charlotte Schaeffer

October 14, 2014

Prompt:

Please outline your educational and professional development plans and career goals. How do you envision graduate school preparing you for a career that allows you to contribute to expanding scientific understanding as well as broadly benefit society? Page limit - 3 pages

Describe your personal, educational and/or professional experiences that motivate your decision to pursue advanced study in science, technology, engineering or mathematics (STEM). Include specific examples of any research and/or professional activities in which you have participated. Present a concise description of the activities, highlight the results and discuss how these activities have prepared you to seek a graduate degree. Specify your role in the activity including the extent to which you worked independently and/or as part of a team. Describe the contributions of your activity to advancing knowledge in STEM fields as well as the potential for broader societal impacts (See Solicitation, Section VI, for more information about Broader Impacts).

NSF Fellows are expected to become globally engaged knowledge experts and leaders who can contribute significantly to research, education, and innovations in science and engineering. The purpose of this statement is to demonstrate your potential to satisfy this requirement. Your ideas and examples do not have to be confined necessarily to the discipline that you have chosen to pursue.

1. Academic Background

Over my years at Miami University, I explored a variety of majors and fields. My first major, declared as a freshman, was mathematics. I enjoyed high-level mathematics and statistics courses, while also exploring a variety of subjects such as economics and organic chemistry. I also took numerous courses in psychology, which unintentionally led to my completion of a second major in psychology. Throughout my exploration of these diverse subjects, I have been most challenged and stimulated by those that required strong analytic thinking and problem solving. For that reason, I continued to take courses in mathematics and statistics even after I had completed the requirements for my major.

I had not had any experience with programming prior to taking a mandatory computer science course at the end of my senior year, in the spring of 2013. I took an introductory class for beginning computer science and engineering students. Taking this class changed the course my academic life. Although I had enjoyed many courses, I had never before felt such a passion for what I was learning. My professor suggested that I take a few more courses in the summer to further evaluate my interest and aptitude, and then to consider applying to graduate school. I took his advice, deferred graduation and took two more computer science courses, Object-Oriented Programing and Data Structures, over the summer. My interest in the subject had only grown. My intention at the end of the summer was to get a minor in computer science, and apply to graduate school in computer science.

I spoke with the Graduate Director about which computer science courses I should be taking to prepare myself for applying to graduate school. He recommended applying to the combined program, as that would make more sense for someone in my particular circumstances. So, I spent the 2013-2014 academic year taking graduate courses in computer science. These courses were challenging, especially those that were more coding intensive. However, I found that I really enjoyed the theory-based courses such as Algorithms and Automata, as they were strongly linked to my background in mathematics. I think that my experience in the Masters of Computer Science program would equate to an accelerated undergraduate program, as I learned a lot of background information in a very short period of time. Additionally, my work for my thesis has given me a lot of experience in computer science research.

2. Thesis Research

For the past year, I have been researching under Dr. John Karro, a professor in the Computer Science Department at Miami University whose main research area is bioinformatics. One significant problem in bioinformatics is the detection of repetitive DNA in a genome. My research seeks to improve the sensitivity of an already existing repeat-finding tool, known as Rapid Ab Initio Detection of Elementary Repeats (RAIDER) [1].

Repetitive DNA is defined as a set of discrete DNA sequences in the same genome that are similar or identical to one another [2]. It makes up a significant portion of most genomes, accounting for over one-third of the genetic material of higher organisms [3]. Therefore, improving the effectiveness of a repeat-finding tool is of great value to the fields of bioinformatics and genomics.

Literature Review

Because the goal of my research is to ultimately improve RAIDER, my first major task was to completely understand RAIDER. I spent the first month or so reading Nathan's thesis and working through it until I had a deep understanding of the approach and the correctness of the algorithm. Additionally, I researched other repeat-finding tools in order to understand how other repeat-finding approaches compare to RAIDER. The idea behind this review is that a thorough understanding of the current state of the art in repeat-finding tools will reveal various approaches on how to best modify and improve RAIDER.

Through this literature review, I gained experience reading, summarizing, and categorizing various sources in order to organize a knowledge base for a particular topic. A doctoral dissertation requires the review of a multitude of background sources, so being able to organize information efficiently is pivotal when seeking a graduate degree.

Tool Evaluation

In order to make any statements about whether my approach improved RAIDER, I needed a way to quantitatively evaluate RAIDER. I worked with Dr. Karro to create an evaluation tool that would compare the results of RAIDER and RepeatScout, a widely used repeat-finding tool, when both were run on the same simulated DNA sequence. This chromosome sequence simulator was created by Dr. Karro. I provided for the necessary steps in order to run RAIDER and RepeatScout on the simulated sequence, using a pipeline to continuously submit jobs to the Miami University cluster and read the results from these submitted jobs. The use of a cluster was necessary in order to quickly obtain the results of these tools.

After getting the results of both RAIDER and RepeatScout, I created a module to calculate relevant statistics about their performance. Since the chromosome being used was simulated, the locations of the repeats in the genome were already known. Therefore, I could determine the number of bases that were accurately and inaccurately categorized as part of a repeat or not part of a repeat (true and false positives or negatives). Such information allowed me to calculate the sensitivity (true positive rate) and specificity (true negative rate) for each tool, in addition to other related calculations.

Creating this evaluation tool gave me experience working with large data sets as well as quantitatively measure the effectiveness of a tool based on its output. Being able to objectively evaluate quality is necessary in order to establish the credibility of a tool, especially when attempting to improve it.

Thesis Research

Repeats can be categorized to be either exact or approximate. An exact repeat is a subsequence that occurs multiple times in the same genome; an approximate repeat is a subsequence that approximately matches multiple other subsequences in the same genome, where two subsequences can differ to some degree and still be considered a valid match. Over long periods of time, originally exactly identical repetitive sequences, which were all copies of some ancestral sequence, have accumulated mutations. Therefore, the detection of approximate repeats can provide information about a particular genome as well as the composition of genomes from previous generations. A repeat-finding algorithm's ability to detect approximate repeats is extremely valuable.

RAIDER, like some other repeat-finding tools, uses spaced seeds to improve the identification of approximate repeats. Spaced seeds are basically patterns describing what positions in two strings must match and what positions in two strings can be a match or a mismatch (don't care positions). RAIDER currently only allows for a single spaced seed to be used in repeat detection, and this seed is chosen arbitarily [1].

Through my thesis research, I hope to improve upon RAIDER's current ability to detect approximate repeats through an in-depth study of spaced seeds. I am currently investigating whether the sensitivity of RAIDER to the detection of approximate repeats could be improved through the use of multiple seeds, more careful seed design and analysis, and/or changes to the algorithm that would make it more amenable to the use of spaced seeds.

3. Graduate Assistantship

My strong mathematical background gave me an advantage when I took theory-based computer science courses. For this reason, I was asked to be the teaching assistant for the Algorithms course professor during my second semester as a graduate student. I was in charge of grading problem sets, as well as holding office hours to provide additional help for students. Due to positive reviews from the professor I was assisting, the Department Head asked me to be the teaching assistant for Computer Architecture. Additionally, I was asked to tutor an undergraduate student in the introductory Software Engineering course.

Upon the suggestion of my research advisor, I applied for and received a Graduate Assistantship for the 2014-2015 academic year. This award covers my tuition and gives me a living stipend in exchange for my work as a teaching assistant for a professor. This semester, I am the teaching assistant for Algorithms, Data Structures, and Operating Systems.

As a teaching assistant, I have experience grading papers, problem sets, and proofs, as well as holding office hours to provide additional help for undergraduate and graduate students in various courses.

4. Amazon Internship

References

[1] N. Figueroa, X. Liu, J. Wang, and J. Karro, "Raider: Rapid ab initio detection of elementary repeats," in *Advances in Bioinformatics and Computational Biology*, pp. 170–180, Springer, 2013.

[2] T. J. Treangen and S. L. Salzberg, "Repetitive dna and next-generation sequencing: computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, pp. 36–46, 01 2012.

[3] R. Britten and D. Kohne, "Repeated sequences in dna," *Science*, vol. 161, pp. 529–540, August 1968.