

Computer Science/Bioinformatics

Proposal to INBRE Spring 2006 course release Holly Patterson-McNeill

Final Report

Proposal: From the undergrad bioinformatics programs that I have researched, the computer science (CS) component tends to be a list of courses without direct relation to bioinformatics. Because CS students are not exposed to the wide world of bioinformatics, there needs to be a way to introduce the topic to them. There needs to be a method to conceptually link bioinformatics data collection, storage, and analysis to computer science tools and techniques. I propose to introduce into three current CS classes, an insight into the use of CS in bioinformatics applications by the use of learning modules that will be developed during the time release granted by this proposal.

Key Objectives:

Develop the introductory computer science/bioinformatics modules for CS111 Foundations of Programming, CS213 Data Structures, and CS 310 Analysis of Algorithms.

The proposed modules for each of the courses listed above have been developed.

1. CS111 Foundations of Programming is an introductory level programming course. The students who take the course are from multiple disciplines, including computer science, math, GIS, engineering, web development, information systems analysis, and justice studies. Because these students lack programming skills, I will be using a case study approach throughout the semester. The problem to be used is the Partial Digest Problem and its Brute Force algorithmic solution. The problem will be introduced at the beginning of the semester. Then as the students gain programming skills, we will work on the solution. This brute force algorithm is easy enough to understand and complex enough to be interesting. It relates concepts that the students will be learning throughout the semester, including binary arithmetic, conditional and repetition statements, and functional decomposition.

In addition to the above module, I will be incorporating bioinformatics into the example code segments that are used to illustrate each new concept. For example, when we study string functions I will use DNA manipulations instead of standard English text. These examples are gleaned from the "Python Course in Bioinformatics", by Katja Schuerer and Catherine Letondal

(<http://www.pasteur.fr/recherche/unites/sis/formation/python/index.html>).

2. CS213 Algorithms and Data Structures is the second course taken by computer science majors. As such, the students are ready for more complex algorithms and structures that can be applied to computational biology. String matching algorithms are suitable for this level student. Although these algorithms are common topics in more advanced courses, I have developed two modules which explore simple approaches to this problem. Instead of the usual 'business' problems to contextualize our programs, we will be using the

scenarios in *Algorithms on Strings, Trees, and Sequences, Computer Science and Computational Biology* by Dan Gusfield as the basis for these programs. I am particularly excited about having the module on trees beyond the traditional binary trees studied in Data Structures. As in the CS111 course, this topic will be spread throughout the semester. When we study lists, then the brute force algorithms are suitable. When we study trees, then keyword tree algorithms provide a new application for these students.

3. CS310 Algorithm Design and Analysis is an advanced course for computer science students. This class is ideal for more advanced algorithms on string matching problems. Traditional algorithms, such as Boyer-Moore and Knuth-Morris-Pratt are typically studied without context. Although this module merely provides context to algorithms already included in the course, framing them within the computational biology context will make their study more interesting. A second module on more advanced string matching algorithms builds on the tree approach used in CS213. Again I use scenarios from *Algorithms on Strings, Trees, and Sequences, Computer Science and Computational Biology* by Dan Gusfield.

All in all, I am very pleased with the new modules developed that tie bioinformatics into the computer science curriculum. The most difficult part of this project has been taking algorithms from advanced textbooks and reworking them into explanations and assignments that are understandable in lower level courses.

Future work: I am trying to find parallel algorithms that are suitable to undergraduate students that are applicable to computational biology. The difficulty here is that the theory is seldom associated with its applications. The parallel algorithms will then be programmed using the Apple Cluster.

Dissemination Plan:

The modules will be made available to the instructors of the three computer science courses at LCSC and the instructors will be highly encouraged to use them. In addition, papers describing the modules and their use will be submitted to the ACM Special Interest Group on Computer Science Education annual conference and the Northwest Consortium for Computer Science in Colleges annual conference.