Using Interdisciplinary Bioinformatics Undergraduate Research to Recruit and Retain Computer Science Students

Jon Beck
Computer Science
Truman State University
Kirksville, MO 63501
jbeck@truman.edu

Brent Buckner
Biology
Truman State University
Kirksville, MO 63501
bbuckner@truman.edu

Olga Nikolova
Bioinformatics and Computational Biology
Iowa State University
Ames, IA 50011
olia@iastate.edu

ABSTRACT

An interdisciplinary undergraduate research project in bioinformatics, jointly mentored by faculty in computer science and biology, has been developed and is being used to provide top-quality instruction to biology and computer science students. This paper explains the benefits of such a collaboration to computer science students and to the computer science discipline. Specific goals of the project include increased recruitment of students into computer science and increased retention within the discipline. The project is also intended to be particularly attractive to women students.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; K.3.2 [**Computers and Education**]: Computer and Information Science Education—*Curriculum*

General Terms

Experimentation, Human Factors

Keywords

bioinformatics, undergraduate research, interdisciplinary research, women in CS

1. INTRODUCTION

This paper describes a program of interdisciplinary collaborative undergraduate bioinformatics research designed to stimulate the interest of computer science students, especially women, to attract new students, to aid in their retention, and to encourage them to enter graduate school in computer science, bioinformatics, or computational biology upon graduation.

The number of undergraduate students and undergraduate degrees granted in computer science is declining in the United States[9]. Furthermore, despite efforts to attract women to the field, the numbers for female students are far below those for males[8]. Various reasons have been put forward for the overall decline, as well as for the gender differential. Among the factors that are causal, or at least associated with, the overall declines, two are particularly relevant to students and faculty at our medium-sized public undergraduate liberal arts institution.

Students don't see the need for a CS degree. Students taking CS courses do not wish to study computers as an end in themselves, but rather to become proficient in their use to the extent that they can use the computer as a tool to accomplish some other, non-computer-related goal. A number of software environments make common programming tasks very accessible to the non-computer scientist. For scientific computation, high-level interpreted environments such as MATLAB®, Octave, and Mathematica® are quite accessible to mathematics, biology, chemistry, and physics majors and do not require formal computer science theory. Similarly, many systems are available for layout, design, publishing, and web page development.

A declining interest in engineering disciplines. We have seen a drop in interest among engineering- and physics-related disciplines in the past 15 years, overall to some degree and among women students in particular. To the extent that computer science is seen as an engineering discipline, it suffers from this phenomenon. Conversely, we have seen an increase in interest in disciplines related to the life sciences.

2. BIOINFORMATICS

Bioinformatics is a relatively new discipline but has rapidly become an important application area of computer science. The Computing Curricula 2001 Report from the ACM lists bioinformatics and computational biology as two of the scientific computing areas that are a "vital part of the discipline," whereas the 1991 document of the same body contains no reference to them[5, 7].

LeBlanc and Dyer argue that computer science is situated to take advantage of the growing importance of bioinformatics in biological education[4]. They found that there is significant overlap and strong mapping between the core knowledge areas contained in the Computing Curricula 2001 Report and the informatics needs of the bioinformatics research community. Further, they argue that computer science can be effectively taught in a setting in which bioin-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGCSE'07, March 7–10, 2007, Covington, Kentucky, USA.
Copyright 2007 ACM 1-59593-361-1/07/0003 ...$5.00.

formatics questions form the rationale and motivation for presenting the core computer science knowledge areas.

The field of biological education is undergoing a major shift, becoming far more quantitative and computational. The National Academies' Bio2010 report clearly delineates this shift, saying that "connections between biology and the other scientific disciplines [chemistry, physics, and mathematical sciences] need to be developed and reinforced so that interdisciplinary thinking and work become second nature"[10]. The report also makes the specific recommendation that all biology students should be encouraged to pursue a research experience "as early as is practical in their education." Bioinformatics is an ideal example of a field that facilitates this desired connection between the life and physical sciences by providing interdisciplinary research experiences.

3. WOMEN IN CS

While bioinformatics is an exciting and inviting applied area for any computer science student, we believe it has special significance for female undergraduates. According to Fisher and Margolis, women and men have different motivations for studying computer science[3]. Specifically, women "link their interest in computer science to other arenas." In our experience, biological investigations are viewed as particularly socially relevant by female students, as they so often have applications that benefit society in areas such as medicine, nutrition, and public health. This, according to Rosser, is an excellent way to attract female students to a science discipline[6, pp 71–72]. A recent study at Point Loma Nazarene University found that the number one reason female students are attracted toward computer science as a major discipline is the desire to use computing in another field[2].

4. OUR PROGRAM

We have embarked on a collaborative interdisciplinary research project involving biology and computer science undergraduate students with faculty mentors. The focus of our research is on the annotation of maize genes and analysis of their expression patterns based on microarray hybridization data. We offer our experiences with this project as a model for effective research-based teaching of undergraduate biology and computer science students and as a means of attracting and retaining CS students in general and women in particular. Our project, partially funded by NSF grants,[1] supports four to six biology students and one to two computer science students annually.

In our project, sophomore year genetics students perform an investigative bioinformatics analysis of microarray data from global gene expression experiments aimed at understanding the shoot apical meristem (SAM) in maize. The SAM is the above-ground undifferentiated tissue at the tip of the growing plant shoot. From this tissue, all above-ground organs (leaves, etc.) arise. An understanding of the gene expression patterns in this highly active primordial region of the shoot is of fundamental importance to biologists. Few of the SAM-specific genes in maize have been characterized to date. In this project, students retrieve DNA sequences from a database and perform BLAST[2] searches on these sequences against the repository databases of GenBank. BLASTN and BLASTX analyses are evaluated in an attempt to provide an annotation, based on nucleic acid sequence identity or amino acid sequence similarity for the expressed sequence tag DNA probe (EST) on the microarray. Then, the students use the EST sequence as a query in the Maize Aligned Gene Islands (MAGI) database[3] to acquire a longer EST which is used as a query in BLASTN and BLASTX analyses. In addition, the students perform InterProScan[4] searches on the longest available sequence in an attempt to identify protein functional domains. These analyses most often provide information that enable the students to annotate the EST with considerable confidence. After each EST is annotated with a gene name, it is placed into a functional category defined to support the ultimate goal of facilitating the understanding of the cellular networks and circuitry involved in SAM function and maintenance, and in leaf primordial formation. This experience challenges the biology students to recall, apply and expand their knowledge gained in courses, especially cell biology, and gives many of them their first opportunity to read the primary literature.

Students in our sophomore-level genetics course perform the above steps manually on a few ESTs in course lab exercises. Biology students for the present research project are recruited into the project from the genetics course, to participate in the annotation of the thousands of ESTs that have been identified as being differentially regulated in the maize SAM.

While all of the above work can be carried out manually by biology students and their faculty mentors with no assistance or intervention by computer scientists, it is an extremely labor-intensive process; so arduous as to make it almost impossible for the undergraduates to get past the mechanics. Making meaningful use of the data set, and adding value to it in the form of the annotations and analyses discussed above, requires a robust data storage and manipulation platform. This platform needs integrated tools for the entire lifecycle of the undergraduate research project. The phases of the project include receiving and loading the initial set of raw microarray expression data, the presentation of this data to the annotators in useful and understandable formats, automated retrieval of auxiliary and associated information such as BLAST searches, review of the students' annotations by the faculty mentors, real-time generation of statistical analyses of the annotation results, and presentation of the results to the scientific community. In addition, the system needs to provide these capabilities securely and simultaneously to multiple authenticated researchers at multiple locations. Finally, the data must be backed up to safeguard the invested effort and time it represents.

The development of such a platform, in our case a relational database-driven system with a sophisticated, dynamic web front end, requires resources far beyond those available to the typical undergraduate biology department. However, it is the perfect vehicle for a research and development project for computer science majors. The development of the system requires significant, intense collaboration and cooperation among biology and computer science students and

[1]NSF award DBI-0321595, Michael Scanlon, PI, and NSF award DUE-0436348, Jason Miller, PI.

[2]http://www.ncbi.nlm.nih.gov/BLAST/

[3]http://magi.plantgenomics.iastate.edu/

[4]http://www.ebi.ac.uk/InterProScan/

their faculty mentors. Beyond the development and maintenance of the system itself, such a collaboration provides multiple and significant opportunities for open-ended spin-off bioinformatics and computer science research projects in such areas as clustering, biological data mining, large data set problems, distributed computing, and software engineering. This interdisciplinary collaboration works to the advantage of both disciplines. It provides computer scientists with a valuable experience in working in an applied field and with training in a highly attractive field at the intersection of biology and informatics. It provides biology students with exposure to informatics concepts. They become quite familiar with the capabilities and shortcomings of informatics approaches to, and automated processing of, biological data.

In our experience, female students have been particularly attracted to the broad interdisciplinary strategies for problem solving inherent in our research collaborative. Rosser advises mentors to "use methods from a variety of fields or interdisciplinary approaches to problem-solving. Because of their interest in relationships and interdependence, female students will be more attracted to science and its methods when they perceive its usefulness in other disciplines"[6, p 64]. One female CS student in our research program stated, "my own experiences ... argue that interdisciplinary research is an extremely effective way to recruit women into computer science research, and to demonstrate the social relevance of computer science as a career discipline for women ... I wish to emphasize the significance of working in interdisciplinary teams formed of undergraduate students and faculty members with different academic backgrounds." Even though this student had no prior experience in biology, she goes on to argue that bioinformatics in particular provided the perfect environment for developing computer science skills and knowledge.

5. UNDERGRADUATE RESEARCH

According to Becker[1], only through research experiences do students "attain a higher level of competence in the science, mathematics, engineering, and technology fields" and in a research setting students best learn and practice the teamwork, complex problem-solving, and communications skills they will require after graduation. Becker goes on to argue, in fact, that *all* computer science undergraduates should engage in research. We agree that research is an essential component of the undergraduate experience and that the learning that takes place in a research setting cannot be duplicated in the classroom.

Our project is therefore designed as a research experience to take advantage of the demonstrated benefits of research for undergraduates; designed as an interdisciplinary bioinformatics endeavor because of the efficacy of bioinformatics to recruit and retain computer science students.

Our undergraduate research project is set within the context of a larger framework. Truman State University has an extensive, mature culture of fostering undergraduate research. The authors are participants in Truman's Mathematical Biology Initiative. This program has the primary goal of increasing the recruitment and production of students equipped to work or enter graduate programs in mathematical or computational biology or bioinformatics. This is done by supporting long-term research opportunities for undergraduate students to analyze problems and to solve them from two complementary disciplinary perspectives, one in biology and one in the mathematical or computational sciences. Participating students are involved in a variety of cross-disciplinary research experiences at the intersection of biology and the mathematical and computational sciences supported by faculty teams. The research projects span the range of scales from the molecular to communities and metapopulations, and include various emphases in mathematical biology, modeling, biostatistics, computational biology, and bioinformatics.

Our research programs have a field trip component in which the undergraduates are taken to Research I graduate schools and to commercial research institutions to meet with faculty and researchers, including their mentors' collaborators. When the undergraduates see the source of their microarray data and present results to the researchers, they gain valuable experience and networking contacts. More importantly, they see the broader relevance of their work and see how it fits into a larger context of ongoing work.

6. COLLATERAL BENEFITS

Establishing a fruitful biology and computer science collaborative requires that the mentors and students from both disciplines make an effort to develop at least a rudimentary understanding of the fundamental principles and methods of each others disciplines. In our experience, mentors and students should commit the time to sit in on introductory courses within the complementary discipline. While this activity is difficult to execute in an otherwise busy schedule it is an excellent opportunity to capture the big picture of a discipline with which one is unfamiliar. There are also other valuable dividends of this commitment. The visiting professor can make relevant comments and contributions during lectures, adding value to the educational experience. Additionally, in our experience, students enrolled in the courses will often enroll in the courses of the collaborator in subsequent semesters. Thus, this collaboration is a valuable mechanism to recruit students to courses. It encourages the cross disciplinary education that is a desired quality for students at a liberal arts institution, especially those considering graduate education. Lastly, students enrolled in the courses readily see the intellectual commitment of the faculty mentors and come to better understand the interdisciplinary nature of modern science.

7. RESULTS

This program is still in its infancy, having run for nearly two years and having had one computer science student complete the program. We believe, however, based on the interest this project has generated among the computer science students here at Truman that collaborative bioinformatics research has the potential to significantly aid our student recruitment and retention. Prior to joining the research project, the female student who has completed the experience was planning to get a job in IT after receiving her bachelor's degree. During the course of her participation in the research project, her interest in bioinformatics developed to the degree that, having completed her baccalaureate, she is now a Ph.D. student in Bioinformatics and Computational Biology at a Research I institution. A second CS student is currently involved in the project and has not yet decided his post-baccalaureate plans.

The project is being used as a recruiting tool by the admissions office and by faculty in visits with prospective students. The women's computer science club, Tru Women in Computer Science (TWiCS), is using the program as a recruiting tool in its outreach and recruiting efforts at area feeder high schools during academic year 2006–2007. The results of these recruiting efforts will be watched carefully to see what effect they have on attracting new students to computer science.

8. ACKNOWLEDGMENTS

The authors recognize the Truman undergraduate biology students for devoting endless hours annotating the genes in this project and for working so closely and willingly with the computer scientists. In particular we thank Kate Browning, Ashleigh Fritz, Lisa Grantham, Eneda Hoxha, Ashley Lough, and Zhian Kamvar, who have made this project so enjoyable. We acknowledge the work of our collaborators in providing the framework for this effort, in particular Michael Scanlon, Patrick Schnable, Marja Timmermans, and Kazuhiro Ohtsu. Many thanks to Kerrin Smith for her valuable contributions to the tone, style, and organization of this manuscript.

9. ADDITIONAL AUTHORS

Additional author: Diane Janick-Buckner, Biology, Truman State University, djb@truman.edu

10. REFERENCES

[1] K. Becker. Cutting-edge research by undergraduates on a shoestring? *J. Comput. Small Coll.*, 21(1):160–168, 2005.

[2] L. Carter. Why students with an apparent aptitude for computer science don't choose to major in computer science. *SIGCSE Bulletin*, 38(1):27–31, 2006.

[3] A. Fisher and J. Margolis. Unlocking the clubhouse: the Carnegie Mellon experience. *SIGCSE Bulletin*, 34(2):79–83, 2002.

[4] M. D. LeBlanc and B. D. Dyer. Bioinformatics and Computing Curricula 2001: why computer science is well positioned in a post-genomic world. *SIGCSE Bulletin*, 36(4):64–68, 2004.

[5] E. Roberts, editor. *Computing Curricula 2001: Computer Science Final Report*. Association for Computing Machinery, 2001.

[6] S. V. Rosser. *Female-friendly science: applying women's studies methods and theories to attract students*. Pergamon Press, Elmsford, NY, 1990.

[7] A. B. Tucker, editor. *Computing Curricula 1991: Report of the ACM/IEEE-CS Joint Curriculum Task Force*. Association for Computing Machinery, 1991.

[8] J. Vegso. Interest in CS as a major drops among incoming freshmen. *Computing Research News*, 17(3), May 2005.

[9] J. Vegso. Drop in CS bachelor's degree production. *Computing Research News*, 18(2), March 2006.

[10] P. T. Whitacre, editor. *Bio2010: Transforming Undergraduate Education for Future Research Biologists*. National Academies Press, Washington, DC, 2003.