AC 2007-1200: A PROJECT-CENTRIC APPROACH FOR CYBERINFRASTRUCTURE IN BIOINFORMATICS

Daphne Rainey, Virginia Bioinformatics Institute

Bruce Mutter, Bluefield State College

Lionel Craddock, Bluefield State College

Susan Faulkner, Virginia Bioinformatics Institute

Frank Hart, Bluefield State College

Martha Eborall, Bluefield State College

Lewis Foster, Bluefield State College

Stephen Cammer, Virginia Bioinformatics Institute

Betsy Tretola, Virginia Tech

Bruno Sobral, Virginia Bioinformatics Institute

Oswald Crasta, Virginia Bioinformatics Institute

© American Society for Engineering Education, 2007

Abstract
Rapid advances in scientific engineering and computer technologies have facilitated the generation of a vast amount of research data. The integration of knowledge from various fields such as computer science, mathematics, chemistry, and biology has resulted in a vast opportunity for creating new research environments based upon cyberinfrastructure (CI). We describe here two projects that were carried out to train the current scientists as well as future workforce to harness the full power of CI for discovery, learning, and innovation across and within all areas of science and engineering. First, the Training Education Advancement and Mentoring (CI-TEAM) demonstration project focused on preparing the future scientific workforce through development and implementation of an interdisciplinary bioinformatics course. Central to the course is a project-centric teaching paradigm to engage students. In this project, the faculty and their students at Bluefield State College (BSC) were introduced to the concepts of CI. The course modules were further modified by BSC to fit the students' and training objectives. We report here the first implementation and assessment of the CI course using BSC's Center for Applied Research and Technology (CART) Course Management Service (CMS). The second project was carried out to involve current scientists through similar project-centric approach using the concepts of CI. The Bioinformatics and Genomics Research Core (BGRC) at VBI, as part of the Mid-Atalantic Regional Center of Excellence (MARCE) provided training and support to over hundred researchers working in the area of emerging infectious diseases to enable them in generation, storage, analysis and/or interpretation of 'omic data. The effective interaction has enhanced discovery of new knowledge as well as feedback for infrastructure development.

1 Introduction

Advances in computational technology are changing the way research is conducted in all aspects of science. Rapid advances in scientific engineering and computer technologies have facilitated the generation of a vast amount of research data. For example, the number of nucleotide sequences in public databases doubles every six months. The integration of knowledge from various fields such as computer science, mathematics, chemistry, and biology has resulted in a vast opportunity for creating new research environments based upon CI (Atkins et al., 2003[1]). In part, this is accomplished by providing effective and efficient platforms that empower scientists and engineers to conduct multi-disciplinary team research.

Bioinformatics is one area where several CI concepts have been successfully implemented through the development of enabling hardware, software, algorithms, and collaborative research support. One of the major challenges facing the post-genomic era is the integration of diverse data sets (Stein, 2003[2]). As described by Kanehisa and Bork (2003) [3], the main goal of bioinformatics during the 1990s was to create primary databases of genes and proteins. Currently, the focus is on extending the databases for quantitative data from transcriptomes and proteomes and providing interoperability among multiple disciplines. The main goal of bioinformatics in the future will be to create a knowledgebase—and the tool set to use the knowledgebase—to advance discovery by implementing the concepts of CI through the integration of various databases, algorithms, and scientific disciplines[4,5].

Our goals were to prepare the current scientists as well as future workforce for the knowledge and requisite skills needed to harness the full power of cyberinfrastructure for discovery, learning, and innovation across and within all areas of science and engineering. The specific aim of the current activities described here were to orient future and current generations of scientists, engineers, and educators to the principles of CI in their teaching, training, and learning. Here we describe two types of CI projects that were developed and implemented. For current scientists engaged in biomedical research, we have developed sustainable training materials through collaborative research projects as part of the Bioinformatics and Genomics Research Core (BGRC). For future scientific workforce, we have developed and implemented an interdisciplinary project-centric bioinformatics course as part of the Cyberinfrastructure Training Education Advancement and Mentoring (CI-TEAM) program.

1.1 CI-TEAM Demonstration Project

In 2005 the Virginia Bioinformatics Institute (VBI) at Virginia Tech University was awarded a CI-TEAM Demonstration project. The CI-TEAM members of this project consists of Bluefield State College (BSC), Bluefield, West Virginia , Galileo Magnet High School(GMHS), Danville Virginia , and VBI. The CI-TEAM Demonstration project began in January 2006. The initial plan for the two-year project was to develop material to create Cyberinfrastructure courses that would be implemented and evaluated by both BSC and GMHS in 2007. VBI's role was to develop course modules based on ongoing projects utilizing cutting edge bioinformatics tools and genomics results to allow for the introduction to students and faculty of each institution to the concepts of CI. The faculty at BSC and GMHS were charged with the task of developing and bringing together materials to supplement the modules and tailor the information to the students at their respective institutions. BSC prepares many non-traditional students for challenging careers, graduate study, informed citizenship, community involvement, and public service in an evolving global society. The college offers undergraduate liberal arts and professional programs in applied sciences, business, education, humanities, social sciences, engineering technologies, and allied health sciences. Central to the course is a project-centric teaching paradigm to engage students in applying the concepts of CI by integrating the disciplines of biology, computer science, mathematics, and statistics through bioinformatics. An important goal was to demonstrate the connections between these often-disparate fields.

In this section, we report here the development and implementation of the first CI course and a summary of our initial observations to aid others in implementation of similar courses. Specifically, we discuss some of the materials that were developed, the use of the CART CMS in course delivery and some of the pedagogical considerations important to course implementation.

1.1.1 Development and Implementation of Project-Centric Cyberinfrastructure Education

Implementation of a project-centric teaching paradigm was aimed at engaging students in applying the concepts of CI. During the process of course development and delivery, we made use of the rapidly increasing volume of biological data and accompanying bioinformatics tools and they served as valuable teaching resources. For the course, real scenarios were used to design projects such that the solutions required contributions from personnel with diverse areas of expertise such as molecular biology, microbiology, and bioinformatics (Figure 1). Accessing databases and analysis tools often requires minimal specialized computer skills (i.e. accessing the Internet); the challenge becomes helping students understand the context of the biological problems that can be addressed with these tools (Greene and Donovan, 2005)[6]. Accordingly, we chose to focus the course on learning modules centered on several key pathogens of interest to biodefense and to public health. The objective was to stimulate interaction and participation and ingrain the CI concept through role playing activities and presentations. Roles were developed around professionals that would come into play should an outbreak occur, such as Center for Disease Control specialists, researchers involved in vaccine and/or drug development, hospital physicians, microbiologists, and evolutionary biologists). Students work in different roles, accessing various data to come to a joint approach for addressing and solving the problem. A focus was placed on the importance of forming transdisciplinary teams and bioinformatics data and tool use. While background was important to the depth of understanding a problem, it was not intended for students to require extensive expertise in any one area.

The key concept of this project is to provide an understanding of CI to students through a problem-based approach rather than a discipline-centric view. The main objective was to stimulate interaction and participation and instill the CI concept through role playing activities and presentations (Rainey et al., 2007) [7]. At BSC, the first cohort of students participated in three learning modules to obtain a sense of integrating biological data through use of CI in realistic scenarios. These scenarios involved known human and animal pathogens in current VBI research. The data were incorporated into three modules focused on analyzing a novel strain of Bacillus anthracis, the causative agent of anthrax; identifying novel drug targets for Rickettsia species, the tick-borne parasites responsible for typhus, Rocky Mountain Spotted Fever and Lyme's disease; and identifying vaccine candidates against Brucella abortus, causative agent for brucellosis in cattle.

During the course the core bioinformatics tools were introduced, focusing on sequence searching and retrieval from National Center for Bioinformatics (NCBI), sequence comparison, sequence manipulation in the form of multiple sequence

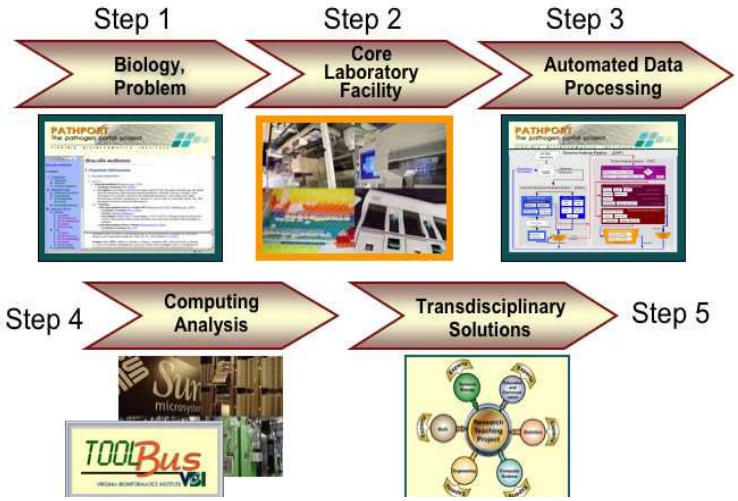

Figure 1. Flow diagram showing development and implementation of a project-centric bioinformatics course for CI-TEAM demonstration.

alignments, phylogenetic tree production, and genome alignment software in Toolbus (He et al., 2005)[8]. A variety of software sources were introduced for sequence alignment and phylogenetic tree production and viewing, however BSC professors tested and chose an appropriate source for its own use.

The Center for Applied Research and Technology, Inc. (CART) Course Management Service (CMS) became the online framework for the course called COSC 490 – Cyberinfrastructure. The CMS software and systems developed by CART at BSC allowed for full online course administration and access to syllabus, outline, surveys, quizzes, testing, reading material, chat, instructor collaboration and general remote student communication.

1.1.2 Assessment and Lessons Learned

During the course of 2006, interaction between faculty and students has deepened our awareness of several key factors that have impacted implementation of the CI course at BSC. A thorough summative assessment was document and observations from faculty and student were assembled in the form of lessons learned.

Pre- and post-course online assessment surveys designed to measure the effectiveness of the proposed activities in the NSF CI-TEAM Demonstration project were administered to the BSC students enrolled in the CI course during fall 2006. Seven students (three females and four males) enrolled in the CI course and all except one completed. The course helped students clarify their goals relative to doing CI research and seeking advanced degrees in preparation for work in a bioinformatics field. Students also confirmed that the course enabled their knowledge of CI, as well as, their ability to effectively discuss CI information with scientists, peers, and professors. Further, students indicated that the course increased their comfort-level in working on multidisciplinary teams. VBI delivered a pre-course training workshop to BSC faculty to enable the design of the course and also provided overall support during implementation of the course in the fall 2006 semester. Surveys of BSC faculty related to the effectiveness of the pre-course training workshop and the overall support also were favorable.

A collection of our observations and recommendations are as follows:
Overview of BSC observations

- The project-centric course design did sufficiently engage students and ensured that a new degree of knowledge was obtained from the different discipline areas outside their respective comfort zones.
- The use of CI platforms, particularly the powerful web-based tools mentioned earlier, did instill a desire within students to utilize similar systems in other cross-disciplinary domains.
- The small number of students in the class recognized a need to further pursue knowledge in the bioinformatics domain, particularly in biology, after having taken the course.
- The professors appeared to remain engaged and interested in pursuing further research in CI and other collaborations with bioinformatics researchers, as a result of having taught the demonstration course.
- Both the students and faculty subsequently expressed an interest in further contributing to the CI at VBI by participating in the internships, design and research in the development of middleware or application components, such as TB/PP visualization plug-ins, new data sources, and new analysis tools.

- Finally, a project-centric course can be designed to serve two or three different student populations and used to disseminate the concepts of CI in an online environment. There remains much work yet to be done in all parts of the course described above.

2 CI Demonstration to Current Scientists Through Collaborative Research and Training

The Virginia Bioinformatics Institute (VBI) serves as the Bioinformatics and Genomics Research Core (BGRC) for the Mid-Atlantic Regional Center Excellence (MARCE) to support countermeasure development research. The BGRC includes VBI's Core Computational Facility (CCF), Core Computational Facility (CCF), and Cyberinfrastructure Group (CIG). The CLF provides services in generation of high-throughput data (HTD) in genomics, transcriptomics and proteomics. The CCF provides a unique bioinformatics computational platform with powerful computers, support databases, applications, data storage and backup. The CIG has developed and deployed information systems for pathosystems biology (PSB), including resources for the curation of genomes and pathosystems, database systems for organizing HTD generated, and software systems for analysis and visualization of the data.

Through the MARCE project, the BGRC has, over the past three years, delivered bioinformatics training courses to infectious disease researchers to promote the use and understanding of bioinformatics tools and algorithms and to establish collaborations with researchers. The MARCE course lecture section reviews the fundamental applications and underlying theory of commonly used analysis algorithms, followed by a hands-on section with the bioinformatics tools system. Since late 2003, we have trained more than 100 researchers representing seven MARCE institutions. An example of the course material prepared and delivered is shown in Figure 2. The data used in training materials were specific to the research area that the scientists were working, while the tools and principles were general. The improved, to a great extent, the interactions and follow up collaborations with the scientists.

Figure 2. Web page showing course material prepared to train current scientists working on developing countermeasures against tularemia disease caused by *Franscisella*.

The BGRC has provided collaborative research support to scientists working on emerging and re-emerging infectious diseases. A total of thirteen different collaborative research projects covering host or pathogen response on pathosystems such as

hemorrhagic fever, anthrax, brucellosis, tularemia, eneteropathogens, and VEE-virus were completed. The data generated has helped towards publication of the results (e.g., Gilchrist et al., 2006)[9]. The collaborations have also helped to integrate the data across several projects and mining for new information that could not be obtained individually.

3 Acknowledgements

This CI-TEAM project was supported by funding from the National Science Foundation (OCI-0537461 to O. Crasta). This BGRC project is funded by NIH/NIAID Cooperative Agreement #U54-AI57168 to B.W. Sobral via subcontract from University of Maryland, Baltimore, M. Levine, P.I.

4 References

1. Atkins, DE, Droegemeier, KK, Feldman, SI, Garcia-Molina, H, Klein, ML, Messerschmitt, DG, Messina, P, Ostriker, JP and Wright, MH. (2003) Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. 3 Feb. 2003 (http://www.communitytechnology.org/nsf_ci_report/)
2. Stein L (2003) Integrating Biological Databases. Nature Reviews (Genetics) 4:337-345.
3. Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. Nat Genet 33 Suppl:305-10.
4. Our cultural commonwealth: The Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences, July 18, 2006
5. Buetow, K (2005) Cyberinfrastructure: empowering a “third way” in biomedical research. Science 308(5723): 821-824.
6. Greene, K. and S., Donovan. (2005) Ramping Up to the Biology Workbench: A Multi-Stage Approach to Bioinformatics Education. Bioscene 31(1): 3-11.
7. Rainey, D., Faulkner, S., Craddock, L., Cammer, S., Tretola, B., Sobral, B.W., and O., Crasta. 2007. A project-centric approach to cyberinfrastructure education. TeraGrid 2007.
8. He, Y., R. R. Vines, A. R. Wattam, G. V. Abramochkin, A. W. Dickerman, J. D. Eckart, B. W. S. Sobral (2004) PIML:the Pathogen Information Markup Language. Bioinformatics 21:116- 121.
9. Gilchrist, CA, E Houpt, N Trapaidze, Z Fei, O Crasta, A Asgharpour, C Evans, S Martino-Catt, DJ. Baba, S Stroup, S Hamano, G Ehrenkaufer, M Okada, U Singh, T Nozaki, BJ. Mann, and WA. Petri, Jr. (2006) Impact of intestinal colonization and invasion on the E. histolytica transcriptome. Molecular and Biochemical Parasitology 147:163-176.