REU Site: Bio-Grid Initiatives for Interdisciplinary Research and Education

Chun-Hsi Huang
Department of Computer Science and Engineering
University of Connecticut, Storrs, CT 06269, USA
huang@engr.uconn.edu

ABSTRACT

The Bio-Grid REU (Research Experience for Undergraduates) Site offers undergraduate students to participate in the research activities associated with the *Bio-Grid Initiatives* conducted at UConn. The initiatives aim at advancing the application of modern computing infrastructures and information technology to research and practice in various life-science disciplines. Training seminars are designed to equip students with preliminary background knowledge such as basic parallel programming skills, large-scale data analytics, and middleware support, *etc.*, as well as some ongoing life-science research projects using these computing methods. Students participate in research activities associated with several collaborative projects supported by a campus-wide computational and data grid. The Site was supported by the national Science Foundation from 08-10 and 12-14.

The REU project introduces such interdisciplinary research work to students in the early stage of their academic career to spark their interest. The project aims at preparing future software engineers to formalize and solve emerging life-science problems, as well as life-science researchers with a strong background in high-performance computing.

Keywords

High-Performance Computing, Computational Biology, Grid and Cloud Computing

1. INTRODUCTION

Research work in life sciences increasingly relies on globally distributed information and knowledge repositories. The quality and performance of future computing and storage infrastructure in support of such research depends heavily on the ability to exploit these repositories, to integrate these resources with local information processing environments in a flexible and intuitive way, and to support information extraction and analysis in a timely and on-demand manner.

For example, there has been an unprecedented accumulation of gene sequence data and data related to the physiology and biochemistry of organisms during the past decade. To date, 343 genomes have been sequenced, and genomes of more than 1,500 organisms are at various levels of completion. This wealth of genomic information dramatically accelerates progress toward a comprehensive understanding of the genetic mechanisms involved in diverse biochemical processes pertinent to bioremediation, medicine, biotechnology and agriculture. Efficiency and accuracy of genetic sequence analysis are achieved by the use of diverse CPU-intensive bioinformatics tools and algorithms (*e.g.*, analysis of global similarities [37], domain and motif analysis [3], analysis of the relevant structural [32] and functional data). Running these tools on the rapidly growing data is a time-consuming process and needs high-throughput computations to get results in a timely fashion. The aggregated (distributed) computational/storage infrastructure of modern grid and cloud infrastructures offers an ideal platform for mining biological information at the extreme scale.

Modern Grid/Cloud technology represents an emerging and expanding instrumentation, computing, information and storage platform that allows geographically distributed resources, which are under distinct control, to be linked together in a transparent fashion [5, 12]. The power lays not only in the aggregate computing ability, data storage, and network bandwidth that can readily be brought to bear on a particular problem, but also on its ease of use. After research efforts in nearly two decades, the grids and clouds have matured and been widely applied to computation-intensive applications, optimized storage of large-scale distributed data, and the intelligent use of these data [7, 11, 13].

UConn Bio-Grid Initiatives

The UConn *Bio-Grid Initiatives* aim at advancing the application of modern information technology and infrastructure to various life-science disciplines. The research activities center on the development and deployment of modern Grid and Cloud technologies toward an infrastructure for automated life-science information integration, extraction and analysis [18, 19, 23]. The computing infrastructure is based on a campus-wide computational and data Grid, an effort initiated in 2004. This infrastructure is supporting several collaborative, interdisciplinary research projects, led by our project staff from Schools of Engineering and Medicine. Also an ongoing effort is the development of general-purpose middleware support for the transfer of sensitive data. Educational programs associated with the Bio-Grid Initiatives include the development of new and re-development of current courses to incorporate inventions coming out of the

EduHPC 2015 November 15–20, 2015, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2831425.2831429

research, a new cross-disciplinary *bio-computing* minor, as well as new degree tracks emphasizing the application of High-Performance Computing to different life-science disciplines. An annual scientific meeting, the *International Bio-Grid Workshop* has been held in conjunction with the research and educational activities and has been serving as a major scientific venue for our research community.

1.1 Project Goals

This Bio-Grid REU Site is a 10-week summer site supporting ten undergrad students annually. Student participants are expected to have a sufficient background in computing or biological science and be interested in both. A series of training seminars, conducted by the project staff, is designed to equip students with preliminary background knowledge. The seminars introduce to students concepts about modern grid and cloud computing infrastructure and related information technology. Ongoing life-science research projects deploying the campus grid are presented to the students. The seminar series also include talks on research projects that are at the planning stage of being grid-enabled.

The Bio-Grid Initiatives are enriched in research, outreach and educational activities. Associated faculty members are enthusiastic about the REU Site to continue offering the opportunity for nation-wide undergraduate students to involve in the various interdisciplinary activities. In addition to the faculty members, graduate students currently on the collaborative projects also participate in mentoring.

The project goals are as follows.

GOAL 1: Promote an interest in the interdisciplinary research work using HPC techniques targeting underrepresented minorities, students with disabilities, women, and veterans from primarily NSF-targeted institutions in the early stages of their academic careers.
GOAL 2: Promote the acquisition of skills allowing computer scientists and biologists to function in an interdisciplinary working/research environment.
GOAL 3: Attract NSF-targeted students to become researchers and professionals in their specialized discipline (computer science/biological sciences) but with the ability for interdisciplinary work.

Institutional support from the University includes current and future assistance from a number of UConn facilities, including the Division of Student Affairs (veteran's support and services, the Office of Multicultural Affairs (recruiting assistance, infrastructure, and workshop implementation), the Office of Undergraduate Research, and the Center for Students with Disabilities. An ethics workshop is coordinated by the Associate Dean of the Graduate School. UConn's commitment to undergraduate research extends throughout the academic year and is described by the Undergraduate Research Office, which was established to provide a broad range of enrichment activities designed to make research available to all undergraduates (not just honors students). The School of Engineering advocates the REU activities, assists in the dissemination of REU posters to targeted groups and supports the operation of an engineering learning center for summer use. In addition to the lab space with individual faculty mentors, the CSE department provides a lab with twenty machines dedicated to the REU.

2. NATURE OF STUDENT ACTIVITIES

The REU Site actively incorporates feedback from students and mentors while structuring the program. Social activities such as cook-outs, pizza parties and baseball games are jointly planned with other REU programs and are scheduled at the organizational meeting before program starts. In general, the program time lines are as follows:

Day 1: Orientation, Campus Tour, Program Overview, Lunch with Faculty, Ice Cream Social
Week 1: Technical Seminars (mornings); Responsible Conduct of Research Seminars (afternoons)
Week 2: Lab Tours, Center (BiBCI: Bioinformatics and Biocomputing Institute) Tours
Week 3: Invited Talks from BECAT (Booth Engineering Center for Advanced Technologies)
Week 4: Seminar on Proposal Writing, targeting NSF Graduate Research Fellowships
Week 5: Remote Bio-Grid Workshop (may be in conjunction with the student Mini-Symposium)
Week 6: Student Proposal Presentations
Week 7: Seminar on Grad School Application
Week 8: Seminar on Technical Writing and Presentations
Week 9: Diversity Forum: Challenges and Choices in Higher Education (Panel Discussions)
Week 10: Bio-Grid Symposium (Final Project Presentations)

A major component of the program is for students to actively involve in one research project. After the training seminars, students discuss with the project staff about their interests and a research supervisor is assigned to each student. Under the supervision of the project staff, students have their lab space and closely interact with their supervisor and graduate students currently on the project. Project supervisors design a component of the assigned project for students to work on. REU students also have the opportunity to remotely attend the annual International Bio-Grid Workshop. Faculty members associated with the Bio-Grid Initiatives are all welcome to propose student projects.

Research Project: Basic Grid Infrastructure
The development of the Bio-Grid portal focuses on establishing an extensible and robust Application Programming Interface (API), based on standard procedures many life-science applications require when executing in a grid environment [14, 16, 17, 20–22, 29]. This involves various compute platforms, using several different native queue managers (*e.g.*, OpenPBS, Condor, fork), a variety of operating systems (*e.g.*, RedHat Linux, IRIX, Windows), and various Wide Area Network (WAN) connections (*e.g.*, GigE, Fast Ethernet, T1) on campus.

The Bio-Grid is based on the Globus Toolkit middleware version 2.2.4 and the web portal is served by Apache HTTP Server version 2.0. (All of the web portal pages are dynamically created using the PHP hypertext preprocessor scripting language, JavaScript, and real-time MySQL database access.) Each web portal page also allows the strict use of security and authentication procedures, defining a fine grained custom interface for each grid user. Several grid user Access Control Levels (ACLs) are defined for un-authenticated, general, system administrator, and grid administrator web portal page access. The base Globus Toolkit middleware Meta computing Directory Service (MDS) information is

stored in the Bio-Grid database and can be queried and displayed directly. The job monitoring system are designed to be an extremely lightweight and non-intrusive tool for monitoring applications and resources on the Bio-Grid.

Student Participation Students on this project are introduced to the basic computing, data infrastructure and administration of the grids and clouds; and potential applications. Project staff guides REU students through the overall infrastructure set-up process and demonstrates the computation job migration over the Bio-Grid via the web-portal. Students also learn to monitor the job status, user statistics, system load status and the bandwidth consumption status over the web-portal. Project supervisors and senior graduate students also work with REU students on proprietary web-portal design.

Research Project: Middleware Development
For a grid or cloud to serve as the infrastructure for on-demand integration of sensitive clinical practice and healthcare information, additional data management and replication mechanisms should be provided, as the current middleware mainly deals with flat files but not to the metadata level. This project concentrates on creating a *Bio Data Management unit* (BDM unit) that interfaces with the middleware and provides additional metadata management capabilities. For example, medical images are stored in DICOM servers [8, 9, 33] in the hospital. Each image may consist of one or several DICOM files representing portions of the imaged body. The first role of the BDM unit is therefore to serve as an interface between the grid Storage Element and the DICOM server by assigning a logical file name to each image. For each new DICOM image generated by an image producer, a logical file name is created and registered into the replication system. There is not necessarily a physical file instance behind this logical file number but rather a virtual file consisting of a set of DICOM files, that can be reconstructed on the fly by the BDM unit if a request for this logic file number comes in. For efficiency reasons, assembled files are cached in a scratch space before leaving.

Student Participation Students on this project are introduced to the inner workings of the middleware, as well as the architecture of the BDM unit. This is a good training project for those with basic operating system background so they know how secure data transmission is completed between sites with heterogeneous administrative systems.

Research Project: Distributed Image Retrieval
This project integrates the BDM and the middleware installed on the Bio-Grid. Take the following test case for example. A cardiologist is looking for medical images similar to a case of his/her patients to confirm the diagnosis. He/She uses a high-level application interface to select the Magnetic Resonance Image (MRI) sequence corresponding to his/her patient and request for similar images. A set of ranked MRI sequences are returned to the physician and he/she can visualize these new cases and read the corresponding diagnosis. This application can be decomposed into a sequence of steps below.

First, the physician enters a query (*e.g.* find the MRI of Mr. X acquired yesterday in this hospital) through the BDM user interface. The BDM unit sends a query for retrieving metadata to the Grid metadata interface. The user authorizations to access the data are checked by the external metadata service and the patient file logical identifier and its associated parameters (imaging modality, region of interest, dynamic sequence, MR acquisition parameters, *etc.*) are returned to the user interface. A request to find all images similar to that of the patient for which a medical diagnosis is known is made. The BDM is then used to route the request to all sites (*e.g.* hospitals) with metadata services. The logical identifiers of all images matching the patient's parameters are returned. A computation to determine level of similarity is conducted. The job submission service of the grid middleware is then used to distribute computations over available working nodes. For each job, the grid replica manager triggers a replication of the input files.

Student Participation Students on this project witness how modern information infrastructure may assist in medical training and practice. The project staff introduces to students the grid-enabled diagnosis expert system. In addition to operating the system, students get to know how life-sciences and modern information technology are bridged. Project supervisors and senior graduate students also work with REU students on the expert system design.

Research Project: Protein Function Studies
The complete sequencing of numerous genomes now brings up the next major challenge in biology to understand how these genes function. By now scientists have only unraveled the functions of a small percentage of the proteins in these genes. Often protein functions ascribe to some recurring sub-structural *motifs* (or *minimotifs*, as they often contain less than 15 amino acids), which usually present specific positional characteristics in the protein sequences [1, 4, 6, 15, 34, 44]. Devising efficient models, computationally or stochastically, to identify potentially *functional motifs* is the first essential step toward realizing protein functions. This project has yielded a web-based program (MnM, *Minimotif Miner*, [2, 39]) to search the proteome database for the presence of minimotifs, computationally, in protein queries. The MnM downloads our minimotif database and several other NCBI databases (RefSeq, LocusLink, HomoloGene, Taxonomy, Pfam, and dbSNP databases) as the input and analyze for potentially functional minimotifs. The MnM program searches proteins for the presence of minimotifs in our minimotif database. A proteome enrichment factor for each minimotif is calculated by dividing the observed number of a minimotif in a proteome divided by the predicted number which is based on its probability and the amino acid frequencies in each proteome. We used the MnM to analyze several proteomes. Statistics show that several minimotifs are enriched in the human proteome [2, 40].

Student Participation Students on this project involve in a few webtools that are currently under development. An ongoing effort is a webtool, powered by an 11-node HPC cluster, for *transcription factor binding site* search and visualization that we call *LASAGNA-Search* (Length-Aware Site Alignment Guided by Nucleotide Association [24–27]). The webtool accepts variable-length TFBSs in addition to PSSMs (Position Specific Scoring Matrices). It currently offers 1,792 precomputed models based on TFBSs and PSSMs collected from the TRANSFAC Public, JASPAR, ORegAnno

and UniPROBE databases. Its search module exploits position dependence for a TFBS-based model whenever performance gain is indicated by cross-validation. Automatic promoter sequence retrieval is supported for seven organisms, which enables visualization of search results in the UCSC Genome Browser. Search results can also be visualized along promoter sequences locally at LASAGNA-Search for any organism. In addition, a Gene Regulatory Network (GRN) can be constructed from search results and visualized locally with various options. LASAGNA-Search allows users to scan promoters for TFBSs without leaving the LASAGNA-Search page. Many features of LASAGNA-Search were designed to be user-friendly. Hence, even without the knowledge of PSSM or TFBS databases and promoter sequence retrieval tools, users can search for binding sites in a promoter sequence and visualize the hits in the UCSC Genome Browser immediately. Students have the chance to get a general understanding of the computational biology and information technology involved in this project.

Research Project: Genomic Knowledge Inference
It is crucial that the massive genomic data produced are well represented so that useful biological information may be efficiently extracted/inferred. A useful tool for effective knowledge representation is the *semantic network* system [30]. A semantic network is a conceptual model for knowledge representation, in which the knowledge entities are represented by nodes, while the edges are the relations between entities [10, 42, 43]. A semantic network is an effective tool, serving as the backbone knowledge representation system for genomic, clinical and medical data. Usually these knowledge bases are stored at locations geographically distributed. This highlights the importance of an efficient distributed semantic network system enabling distributed knowledge integration and inferences. The semantic network is a key component of the *Unified Medical Language System* (UMLS) project initiated in 1986 by the U.S. National Library of Medicine (NLM). The goal of the UMLS is to facilitate associative retrieval and integration of biomedical information so researchers and health professionals can use such information from different (readable) sources [31, 35]. Our research team has developed a distributed semantic network system, based on a task-based and message-driven model to exploit both task and data parallelism while processing queries.

Student Participation Students learn about semantic networks, the UMLS, biomedical knowledge representation and basic concepts of distributed knowledge reasoning. Students also involve in the design of the distributed UMLS. The design experience of the distributed UMLS, particularly the task model for cooperative inference and the layered architecture for the host and slave systems, complies with the bio data management (BDM) unit, in which the metadata management capabilities can easily be designed. The BDM design for secure retrieval of sensitive bio data was motivated by our research on a distributed UMLS. Students also learn the system design of the distributed UMLS.

Bio-Grid Workshop Participation
We initiated in 2003 the Annual *Bio-Grid* Workshop (International Workshop on Biomedical Computations on the Grid). The aims of the Bio-Grid workshop are to promote and reinforce awareness of the possibilities and advantages linked to the development, deployment and evaluation of Grid technologies in broadly biology-related research and practice. Here the context of "biology" covers the whole range of information from molecular level (genetic and proteomic information) through cells and tissues, to the individual (clinical practice) and the population level (social healthcare). Since its initiation, BioGrid events have usually been held world wide in June or July in such cities as Tokyo, Chicago, Cardiff (UK), Singapore, Rio de Janeiro, and Berlin (Germany), all featuring invited speeches, a technical program and plenary sessions.

3. THE RESEARCH ENVIRONMENT

The university has a *104-node HPC Cluster* and a *64-node SGI Altix 3700* available for summer students. The 104-node (1,408-core) HPC cluster contains both CPU and GPU resources, enabling a massive amount of computing power. It is There are 3 classes of nodes available in the cluster: (1) *60 Intel Westmere compute nodes*, each node with 12 Intel Xeon X5650 Westmere cores (total of 720 cores) and 48 GB of RAM. (2) *40 Intel Sandy Bridge compute nodes*, each node with 16 Intel Xeon E5-2650 Sandy Bridge cores (total of 640 cores) and 64 GB of RAM; and (3) *4 NVIDIA GPU nodes*, each node with 8 NVIDIA Tesla M2050 GPUs, 12 Intel Xeon X5650 Westmere cores and 48 GB of RAM. The nodes are configured to use LSF job scheduler, and uses uses low-latency Infiniband network interconnect. 200 TB of high-performance GPFS storage is made available to the cluster. The SGI Altix system consists of (1) an SGI Altix 350 system with 8 Intel Itanium 2 Processors (1.5GHz with 6MB cache), each with 8 GB Memory, and an 80 GB SATA Hard Drive as the front-end server; and (2) an SGI Altix 3700 Bx2 system with 64 Intel Itanium 2 Processors (1.5GHz with 4MB cache), each with 64 GB memory and 146 GB 10K RPM SCSI Hard Drive, with additional four 300 GB 10K RPM Ultra320 SCSI Hard Drives. These twin systems are networked to the School's existing SGI Onyx-4 Visualization System to provide a complete compute-to-visualization package for project staff and summer students.

The project staff has extensive experience in undergrad teaching, research supervision, and curriculum development. The REU staff and faculty affiliated with Bio-Grid Initiatives have long collaborated on interdisciplinary research projects and have been involving undergraduate students in the research work. The CSE faculty represents a wide diversity of research disciplines and are recipients of major research grants from federal agencies such as the NSF, NIH, DARPA, ONR, *etc.* The twenty-one faculty members came from eleven different countries, representing another dimension of diversity. The department has a long tradition of being at the forefront of computing education, offering broad based educational programs in computer Science and Engineering with bachelor, master and doctoral degrees. The undergraduate degree program was accredited by ABET's Engineering Accreditation Commission (EAC) and Computing Accreditation Commission (CAC), remaining one of a very few programs in US to receive dual ABET accreditation.

4. RECRUITMENT AND SELECTION

In addition to a dedicated REU website linked from the NSF, we encourage student applicants from across the nation by extensively sending out brochures and invitations.

Applicants from institutions where research opportunities are not available are targeted especially. We have previously collaborated with investigators from minority institutions on NSF funded research and educational projects. UConn has a rich history of working closely with minority institutions. We work closely with the university to publicize the REU site among those traditionally minority institutions[1] and among student members of the *National Society of Black Engineers*, the *Society for Hispanic Professional Engineers*, and the *Society of Women Engineers* in the nation through emails, phone contacts, posters and brochures. We work closely with the *National Center for Women & Information Technology* (NCWIT) to actively promote and advocate the ten-week summer REU program to women. Through the academic alliance, NCWIT assists in reaching out across the nation for the purposes of recruitment of women to participate in the Bio-Grid summer REU program. The NCWIT academic alliance includes such NSF targets as minority-serving colleges and institutions lacking research opportunities.

We also contact collaborators and administrative staff at minority institutions to distribute the REU Site posters. Every effort is made to encourage undergraduate student participation from these institutions. Due consideration is given to minority applicants. At the state level, we host workshops at the Connecticut State University System (Central, Southern, Eastern and Western) and the CT Community College System to advertise the REU program, especially minority institutions in CT such as the *Capital Community College* (Hartford) and the *Housatonic Community College* (Bridgeport), *etc.* We work closely with the University to encourage underrepresented students registered at the University of Connecticut to participate in the planned activities. Local students may participate in selected activities at their interests in a flexible manner. They are counted in the quota of 10 students to be accommodated by this REU site. Local students may also participate by registering an independent study session for credits with the faculty supervisor whose project they feel interested in.

We work with the *Office of Veterans Affairs and Military Programs* at UConn to actively recruit veteran students via the *National Association of Veterans' Programs Administrators* (NAVPA). NAVPA's member institutions include such NSF targets as minority-serving community colleges and other institutions lacking research opportunities. The primary role of the office at the university is to provide direct support for student veterans as well as to create a university-wide support network for veterans throughout the university. Additionally, the Division of Student Affairs at UConn owns a house that serves as a drop in center for student veterans. This house has been available to REU student veterans. The office also assists in accessing student veterans across the nation for recruitment.

We work with the *Center for Students with Disabilities* (CSD) at UConn to actively recruit disabled students from UConn and other institutions to participate in the summer REU. UConn's CSD has seen the rapid growth in the number of students with disabilities accessing higher education. In 1993, the University served approximately two hundred students with disabilities. In 2008-2009, the CSD worked with more than one thousand. The CSD program has been recognized as a "model program" by the *Association on Higher Education and Disability* (AHEAD). The CSD has an extensive peer mentor network where entering students with disabilities are paired with current students to assist with their transition to the university community. The service is extended to summer REU students with disabilities. The CSD is also committed to offering specialized training required for mentors to accommodate REU students.

The application requires (1) personal information including prior research, project and course experiences; (2) a brief research statement about personal and professional goals and how the REU experience would further these goals; (3) the selection and prioritization of three potential research topics; and (4) provide two letters of reference, at least one of which should be from the applicant's institution.

Our recruitment goals are to attract a group of highly motivated students, roughly half majoring in computer science and half majoring in a biology-related field, with appropriate academic records and a targeted percentage from non-PhD granting institutions. However, due to the partnerships we have established with former participants' schools and other institutions, we have been using our contacts at each to help select promising students whose academic background may miss the bar yet show high potential in non-GPA ways. Participating faculty of the REU project performs preliminary evaluations of applications, based on academic performance, research statement, and letters of reference. Final decisions are based on the student's (a) academic background and perceived interest and enthusiasm for the interdisciplinary research, (b) home institution (non-PhD or limited research given preference), (c) ethnic and gender diversity, if available (applicant's gender, race and ethnic background are requested on an optional basis), and (d) disability and veteran status. A modified rolling admission procedure is employed until all slots are filled. Mentors and students are matched based on mutual preferences.

5. EVALUATION AND REPORTING

Evaluation of the student experience commences with an initial survey in which REU students articulate their goals for the program. Formative evaluation, *i.e.*, continued evaluation with the possibility of program modification as indicated, is conducted during the first two-three weeks and at the mid-point in addition to summative evaluation (end-of-program surveys for both students and mentors, and follow-up communication with students). The project applies both qualitative and quantitative evaluation methods to assess the goals, as well as the outcome-based objectives for each goal. Assessment is aimed at measuring objectives/outcomes associated with students' acquisition of research skills, especially in the appropriate cross-discipline, gaining an understanding of life in the research lab, and eventual enrollment in appropriate graduate programs. To assess success of the project in recruiting diverse REU students, we track all applications, including demographic data that is supplied by the applicant. We determine what level of recruitment (personal letter, phone call, recruitment visit, *etc.*) provides the best return - both in attracting applicants and in attracting participants - for our recruitment effort. Both qualitative (individual and group interview) and quantitative (survey) approaches are used to assess student experiences and outcomes, and focus on the specific objectives.

Upon arrival, REU students are given a questionnaire on

[1]United States Department of Education Accredited Postsecondary Minority Institutions.

their goals and expectations. Each lab is visited at least once/week during the first month of the program to informally monitor progress as part of ongoing formative evaluation. Interview questions and observations are aimed at determining how well the REU student "fits" in the lab, including the level of interaction with other lab personnel and the student's perceptions regarding how s/he is progressing on a project. We plan a formal mid-program evaluation, possibly in the context of a group breakfast or lunch meeting. Mentors (REU faculty) are also surveyed (a) within the first few weeks; (b) at midpoint and (c) at the end of the program to determine their perceptions of each student's progress. Students' ability to communicate results are assessed both by mentor's observations and by the final presentation which culminate the summer REU experience. At the end of the summer, REU students are given comprehensive survey questionnaires designed to elicit feedback on all aspects of the program, from housing and social activities to a detailed evaluation of their research experience and their intellectual experiences in the form of seminars and ancillary activities. This mixed method designed for assessment can, especially after the program is in place for a few years, provide data to assess the efficacy of our REU site at promoting increased representation of underrepresented minorities, veterans and persons with disabilities in research.

Post-program contact with the students is facilitated by the mentor-student relationships developed over the summer program, in addition to utilizing traditional phone numbers and email addresses. We encourage students to take home their posters from the end-of-summer conference, and contact a faculty member at each student's home institution to let them know about the poster and encourage them to have the REU student give a talk. Students are contacted by email periodically during the following semesters and polled as to their plans and career choices. We solicit the first year REU students as ambassadors and ask them to encourage future applicants from their home colleges.

6. PROJECT OUTCOME

The Bio-Grid REU Site was first funded by NSF CCF-0755373 for 10 students each summer from 08-10. We brought a total of 30 students through the three summer programs (29 completed the program, one student left the 2010 program prior to completion due to family issues) The subsequent renewal was funded by NSF OCI-1156837.

Students conducted research under the supervision of mentors, interacting with graduate students and postdocs, participating in research meetings and experienced life as a researcher. Whenever possible, a Bio student and a CS student were teamed up to experience interdisciplinary teamwork. Students overwhelmingly reported that the REU gave them the understanding they needed to assess whether a research career is for them. Ongoing formative evaluation, consisting of meetings with students at the end of the first week as well as mid-experience, allowed for timely corrections of any problems. Mismatches between student and mentor were resolved early in the program, allowing these students to successfully complete the project. All mentors reported that students were good team members. All students reported viewing research as largely a team effort, and gave specific examples of being able to ask questions of different grad students, or of learning different techniques from different groups. All mentors reported students made progress toward independence in research, with a majority reporting students were able to work with complete independence by the end of the program.

Technical seminars in the first week of program invited mentors to give a research presentation. This provided sufficient project and mentor information to students before the project assignment at the end of the first week. Departmental seminars on the *graduate school application process*, *research proposal writing* targeting the NSF graduate fellowship application, *technical report writing* and *effective scientific presentation* have also been very well received. *Weekly group meetings* with the project staff helped resolve any potential problems in a timely manner. *Joint REU seminars* were integral to our program. Among them are an *Ethics workshop* presentation addressing ethics in research, a *stem cell ethics seminar*, and a discussion regarding ethics. Starting 2010, the ethics workshop evolved into an RCR (*Responsible Conduct of Research*) seminar series, spanning the entire first week of program. The *Diversity in the Sciences Workshop* evolved from a presentation to a panel discussion, utilizing faculty and staff from across the campus. The mid-term *Bio-Grid mini-Symposium* required all Bio-Grid students to prepare a 15-min or so oral presentation, accompanied by a one-page research summary. The mini-symposium was designed to get students familiar with the "proposal defense" stage when in graduate school. It also equipped students with knowledge about grant proposal writing. The final Bio-Grid Symposium program was the culminating event. Each project was alloted 30-min for an oral presentation. A technical report was due before students left the program. Collegiality was fostered by many social events - most of them involving all summer research participants (ice cream social, barbecue, baseball game) and others just the Bio-Grid REU program (initial orientation). Summer research students were housed together in a separate section of the summer dormitory, and interacted with Community Assistants (grad students working in Housing) who were connected to the summer research program.

The first goal was met. The objective of recruiting half CS students and half Bio-related students was being approached, with an improvement from 8:2 (CS:Bio) in 2008 to 6:4 (CS:Bio) in 2010. (Bioinformatics students are put in the Bio category.) The ratio slightly dropped in 2012 to 4:2 (CS:Bio) and in 2013 to 7:2 (CS:Bio). Although the extent to which goals (2) and (3) are successfully accomplished can only be determined with the passage of time, the structured interviews during the course of the summer program and through follow-up interviews immediately after the completion of the program has a strong indication of the successful accomplishment of these goals. It's also noted that most students indicated there was no direct cause-effect relationship between the REU experience and their career choice, although all students acknowledged the interdisciplinary research experience and agreed it has provided valuable input in their post-program research and course activities.

We regularly contact Bio-Grid alumni and offer assistance in recommendation letters and in preparation for graduate fellowship applications for those applying to graduate school. As of July 2015, Bio-Grid alumni have won a 2012 NSF Graduate Research Fellowship, a 2012 Alfred P. Sloan Foundation Graduate Scholarship, an honorable mention of 2012 NSF Graduate Research Fellowship, and a 2010 GAANN Fellowship, US Dept. of Education.

Among the 15 REU students recruited from 12-14, 8 are now in graduate school (6 PhD, 2 MS). 4 students are in the process applying to graduate school. By summer 16 around 80% of REU students recruited from 12-13 will be pursuing a graduate degree. Among the 29 students recruited from 08-10, 13 students are attending or have finished graduate school (7 PhD, 1 MD/PhD, 1 MD, 4 MS). 6 students currently in industry are applying to graduate school and expect to start an MS/PhD program within a year. One student is applying to medical school. In summary, by next summer around 70% of REU students recruited from 08-10 will have completed (or be pursuing) a graduate degree.

Additional student outcomes include peer-reviewed scientific publications co-authored with summer REU students in journals, conferences and workshops in both HPC and bioinformatics [28,36,38,41,45–48]. The outcome of the REU program also involves the benefits faculty and the CSE department would gain from the investment of time and resources in mentoring REU students. Experience gained from our REU Site program concerning assessment and recruitment was used extensively in a few current and pending grants by colleagues. Throughout the funding periods, nine faculty members and fifteen graduate students have mentored summer students.

7. REFERENCES

[1] A. Apostolico and G. Bejerano. Optimal Amnesic Probabilistic Automata or How to Learn and Classify Proteins in Linear Time and Space. In *Proceedings of Fourth International Conference on Computational Molecular Biology (RECOMB)*, pages 25–32, 2000.

[2] S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C.-H. Huang, S. Rajasekaran, J. del Campo, J. Shinn, W. Mohler, M. Maciejewski, M. Gryk, B. Piccirillo, S. Schiller, and M. Schiller. Minimotif Miner: A New Tool for Investigating Protein Function. *Nature Methods*, 3(3):1–3, 2005.

[3] A. Bateman et al. The Pfam protein families database. *Nucleic Acids Res.*, 30:276–280, 2002.

[4] G. Bejerano and G. Yona. Modeling Protein Families Using Probabilistic Suffix Trees. In *Proceedings of Third International Conference on Computational Molecular Biology (RECOMB)*, pages 15–24, 1999.

[5] F. Berman, G. Fox, and T. Hey. Grid Computing: Making the Global Infrastructure a Reality. *John Wiley & Sons*, 2003.

[6] E. Birney. Hidden Markov Models in Biological Sequence Analysis. In *IBM J. RES. & DEV 45(3/4)*, pages 449–454, 2001.

[7] R. Butler, D. Engert, I. Foster, C. Kesselman, S. Tuecke, J. Volmer, and V. Welch. A National-Scale Authentication Infrastructure. *IEEE Transactions on Computer*, 33(12):60–66, 2000.

[8] D. Collins, J. Montagnat, A. Zijdenbos, and A. Evans. Automated Estimation of Brain Volume in Multiple Sclerosis with BICCR. *Information Processing in Medical Imaging*, 2001.

[9] G. Comi, M. Philippi, V. Martinelli, G. Sirabian, A. Visciani, A. Cambi, S. Mammi, M. Rovaris, and M. Canal. Brain Magnetic Resonance Imaging Correlates of Cognitive Impairment in Multiple Sclerosis. *Journal of Neurological Science*, 115:66–73, 1993.

[10] M. P. Evett, J. A. Hendler, and L. Spector. Parallel Knowledge Representation on the Connection Machine. *Journal of Parallel and Distributed Computing*, 22:168–184, 1991.

[11] I. Foster. The Grid: A New Infrastructure for 21st Century. *Physics Today*, 55(2):42–47, 2002.

[12] I. Foster and C. Kesselman. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco, 1999.

[13] M. L. Green and R. Miller. Molecular Structure Determination on a Computational and Data Grid. In *Proceedings 4-th IEEE/ACM Symposium on Cluter Computing and the Grid - BioGrid Workshop, CD-ROM*, 2004.

[14] X. He and C.-H. Huang. Communication Efficient BSP Algorithm for All Nearest Smaller Values Problem. *Journal of Parallel and Distributed Computing*, 61:1425–1438, 2001.

[15] S. Henikoff and J. G. Henikoff. Amino acid Substitution Matrices From Protein Blocks. In *Proceedings of Natl. Acad. Sci., 89*, pages 10915–10919, 1992.

[16] C.-H. Huang. Grid-Enabled Parallel Divide-and-Conquer – Theory and Practice. In *Proceedings of the 17th ACM Symposium on Applied Computing, Madrid, Spain*, pages 865–869, 2002.

[17] C.-H. Huang. Parallel Pattern Identification in Biological Sequences on Clusters. In *Proceedings of the 4th IEEE International Conference on Cluster Computing (IEEE Cluster)*, pages 127–134, 2002.

[18] C.-H. Huang. Bio-Grid: A Collaborative Environment for Life-Science Research. In *Proceedings of the 20-th International Symposium on Critical Care and Medicine*, pages 123–132, 2005.

[19] C.-H. Huang. Bio-Grid: Bridging Life Science and Information Technology. In *Proceedings of the 5-th IEEE/ACM Symposium on Cluster Computing and the Grid (BioGrid Workshop), CD-ROM*, 2005.

[20] C.-H. Huang and X. He. Communication-Efficient Bulk Synchronous Parallel Algorithm for Parentheses Matching. In *Proceedings of the 10th SIAM Conference on Parallel Processing for Scientific Computing, Portsmouth, VA. unpaginated, 9 pages*, 2001.

[21] C.-H. Huang and X. He. Finding Hamiltonian Paths in Tournaments on Clusters – A Provably Communication-Efficient Approach. In *Proceedings of the 16th ACM Symposium on Applied Computing, Las Vegas*, pages 549–553, 2001.

[22] C.-H. Huang and X. He. Parallel Range Searching in Large Databases Based on General Parallel Prefix Computation. In *Proceedings of the 10th SIAM Conference on Parallel Processing for Scientific Computing, Portsmouth, VA. unpaginated, 3 pages*, 2001.

[23] C.-H. Huang and S. Rajasekaran. High-Performance Parallel Biocomputing. *Parallel Computing Journal*, 30(9-10):999–1000, 2004.

[24] C. Lee, A. Abdool, and C.-H. Huang. Pca-based population structure inference with generic clustering algorithms. *BMC bioinformatics*, 10(Suppl 1):S73, 2009.

[25] C. Lee and C.-H. Huang. Searching for transcription factor binding sites in vector spaces. *BMC bioinformatics*, 13(1):215, 2012.

[26] C. Lee and C.-H. Huang. Lasagna: A novel algorithm for transcription factor binding site alignment. *BMC bioinformatics*, 14(1):108, 2013.

[27] C. Lee and C.-H. Huang. Lasagna-search 2.0: integrated transcription factor binding site search and visualization in a browser. *Bioinformatics*, page btu115, 2014.

[28] C. Lee, B. Nkounkou, and C.-H. Huang. Comparison of lda and sprt on clinical dataset classifications. *Biomedical informatics insights*, 4:1, 2011.

[29] C.-W. Lee and C.-H. Huang. Toward Cooperative Genomic Knowledge Inference. *Parallel Computing Journal*, 30(9-10):1127–1135, 2004.

[30] C.-W. Lee, C.-H. Huang, and S. Rajasekaran. TROJAN: A Scalable Parallel Semantic Network System. In *Proceedings of the 15th IEEE International Conference on Tools eith Artificial Intelligence*, pages 219–223, 2003.

[31] D. Lindberg, B. Humphreys, and A. McCray. The Unified Medical Language System. *Methods Inf. Med.*, 32(4):281–291, 1993.

[32] L. LoConte, S. Brenner, T. Hubbard, C. Chothia, and A. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, 30:264–267, 2002.

[33] N. Losseff, L. Wang, H. Lai, D. Yoo, M. Gawne-Caine, W. McDonald, D. Miller, and A. Thomas. Progressive Cerebral Atrophy in Multiple Sclerosis: A serial MRI study. *Brain*, 119(6):2009–2019, 1996.

[34] H. M. Martinez. An Efficient Method for Finding Repeats in Molecular Sequences. *Nucleic Acids Research 11(13)*, pages 4629–4634, 1983.

[35] A. McCray, S. Srinivasan, and A. Browne. Lexical Methods for Managing Variation in Biomedical Terminologies. In *Proceedings Annual Symposium Compu. Appl. Med. Care*, pages 235–239, 1994.

[36] B. Nkounkou, C. Lee, C.-H. Huang, and C. Brown. Biological data classifications with lda and sprt. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 164–168. IEEE, 2010.

[37] W. Pearson. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.*, 24:307–331, 1994.

[38] S. Quader, N. Snyder, K. Su, E. Mochan, and C.-H. Huang. Ml-consensus: a general consensus model for variable-length transcription factor binding sites. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 25–36. Springer, 2011.

[39] S. Rajasekaran, S. Balla, C.-H. Huang, V. Thapar, and M. Schiller. Exact Algorithms for Motif Search. *Journal of Clinical Monitoring and Computing*, 19(4).

[40] S. Rajasekaran and C.-H. Huang. A Randomized Algorithm for Distance Matrix Calculations in Multiple Sequence Alignment. In *Proceedings of First Knowledge Explorration in Life Science Informatics (Kelsi), LNAI 3303, Springer-Verlag*, pages 33–45, 2004.

[41] D. Sharma, S. Balla, S. Rajasekaran, and N. DiGirolamo. Degenerate primer selection algorithms. In *Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB'09. IEEE Symposium on*, pages 155–162. IEEE, 2009.

[42] K. Stoffel, J. Hendler, J. Saltz, and B. Anderson. Parka on MIMD-Supercomputers. Technical Report CS-TR-3672, Computer Science Dept., UM Institute for Advanced Computer Studies, University of Maryland, College Park, 1996.

[43] M. Surdeanu, D. I. Moldovan, and S. M. Harabagiu. Performance Analysis of a Distributed Question/Answering System. *IEEE Trans. on Parallel and Distributed Systems*, 13(6):579–596, 2002.

[44] R. L. Tatusov, Altschul, S. F., and E. V. Koonin. Detection of Conserved Segments in Proteins: Iterative Scanning of Sequence Databases with Alignment Block. In *Proceedings of Natl. Acad. Sci., 91*, pages 12091–12095, 1994.

[45] N. T. L. Tran, L. DeLuccia, A. F. McDonald, and C.-H. Huang. Cross-disciplinary detection and analysis of network motifs. *Bioinformatics and Biology insights*, 9:49, 2015.

[46] N. T. L. Tran, S. Mohan, Z. Xu, and C.-H. Huang. Current innovations and future challenges of network motif detection. *Briefings in Bioinformatics*, 16(3):497–525, 2015.

[47] C. Wong, Y. Li, C. Lee, and C.-H. Huang. Ensemble learning algorithms for classification of mtdna into haplogroups. *Briefings in bioinformatics*, 12(1):1–9, 2011.

[48] E. Wong, B. Baur, S. Quader, and C.-H. Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2012.