Research Experience for Undergraduates

- Funding provided by National Science Foundation.
- Usually 10 paid positions for (junior, senior) students for a 10-week (or so) summer program.
- Housing and dining are available for students.
- Students also receive stipends and travel reimbursement.
Program Timeline

Major Events

- 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 Tours
- Week 3: Invited Talks
- 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: Bio-Grid Mini-Symposium (Student Proposal Presentations)
- Week 7: Departmental Seminar on Grad School Application
- Week 8: Departmental 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)
Example Projects

- Basic Grid and Cloud Infrastructure
- Middleware Development
- Distributed Biomedical Image Retrieval
- Protein Function Studies
- Genomic Knowledge Inference
- Others
The Bio-Grid Initiatives (1)

Campus-Wide Computational Grid

- Base Grid Toolkit middleware installed.
- Provide a robust and extensible grid-enabling APIs.
- Provide a (web) contact portal.
- Define Access Control Levels for grid users. (e.g. Unauthenticated, General, System admin., Grid admin., etc.)
- Provide a job monitoring system.
- Aggregate compute platform statistics.

EduHPC 2015, Austin TX
The Bio-Grid Initiatives (2)

Campus-Wide Data Grid

- Base Grid Toolkit middleware installed.
- Provide a robust and extensible grid-enabling APIs.
- Provide a (web) contact portal.
- Define access control levels for grid users.
- Provide basic file management functions.
- Support transparent data migration between resources (file migration is a function of global/local file aging and resource usage)
- Continually evolve to minimize network traffic and maximize disk space utilization on a per user basis.

EduHPC 2015, Austin TX
Associated Activities (1)

- Ethics and Diversity in Research
Associated Activities (2)
- Basic Research Methodology

Project Summary

- This research project investigates the information technology and infrastructure to better support automated and on-demand knowledge inference from geographically distributed biomedical databases. Specifically, the project develops efficient techniques to extract and analyze knowledge from a distributed Unified Medical Language System (UMLS). The underlying information infrastructure is a campus-wide computational and data Grid. Related research work is part of the Bio-Grid research initiatives currently conducted at the University of Connecticut.

- This research work involves a few tasks, including (1) the development of a distributed semantic network system, based on a task-based and message-driven model to exploit both task and data parallelism while processing queries; (2) the parallelization of the inference engine to speed-up the query processing; and (3) automated data migration among the distributed knowledge bases to maximize the storage utilization rate. The current information infrastructure, as a test-bed, of this project is a campus-wide computational and data Grid. Participating sites of this infrastructure include the Schools of Engineering, Public Health and Medicine at the University of Connecticut. Note that the Grid represents a rapidly emerging and expanding technology that allows geographically distributed resources (CPU cycles, data storage, sensors, visualization devices, and a wide variety of internet-ready instruments), which are under distinct control, to be linked together in a transparent fashion. The aggregate computing power, data storage, network bandwidth, as well as the user friendliness have rendered the Grid a prosperous infrastructure in support of automated processing of distributed information. Our system features multi-threading and task migration to support communication latency hiding and load balancing, respectively.

- Grid-enabling the UMLS is part of the Bio-Grid initiatives currently conducted at the University of Connecticut. The initiatives intend to establish a campus-wide information infrastructure in support of automated and on-demand distributed knowledge exploration in biomedical informatics, which is increasingly needed by life-science research work. Application areas in biomedicine include the epidemiological studies and medical imaging, which produce tremendous amount of data that are usually geographically distributed among hospitals, clinics, research labs, and radiology centers, etc. For research, training or clinical purposes, physicians and researchers often need to consult and analyze data from distributed sites. Thus, an infrastructure supporting on-demand and automated information extraction and reasoning will provide significant convenience.

Components for Summer Students

- Students participating in this project 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 health data management (HDM) unit, a core component under development, in which the metadata management capabilities can easily be designed. The HDM design for secure retrieval of sensitive bio/medical/health data was motivated by our research on a distributed UMLS. Students also learn the overall system design concept of the distributed UMLS.
Classification Techniques for mtDNA SNPs into Haplogroups

Li, Yuran1,3,* Wong, Carol2,4,*

2009 National Science Foundation – BioGRID R1U Fellows
Department of Computer Science and Engineering
University of Connecticut, Storrs, CT 06269

1Department of Chemistry and Biochemistry, University of Delaware, Newark, DE 19717
2Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104
3yl4@udel.edu 4cwongh@seas.upenn.edu

*These authors contributed equally to this work

Abstract. The classification of mitochondrial DNA (mtDNA) into their respective haplogroups allows the addressing of various anthropologic and forensic issues. Unique to mtDNA is its abundance and uncombining uniparental mode of inheritance, consequently, mutations are the only changes observed in the genetic material. These individual mutations are classified into their distinct haplogroups allowing the tracing of different genetic branch points in human (and other organism) evolution. The Genographic Project recently began assembling a vast database of mtDNA categorized into their respective haplogroups. Due to the large number of samples (~79,500) present, it is necessary to automate the classification of mtDNA into haplogroups using various computer algorithms as well as evaluate the accuracy of each algorithm. Using a five-fold cross validation test, the nearest-neighbor method employed by the Genographic Project achieved a macro accuracy rate of 87.37% and micro accuracy of 96.26% using a consented database of 21,141 samples. Two more state-of-the-art classification techniques, support vector machines and random forest, were employed and evaluated on the same database and their accuracy also evaluated using five fold cross validations. The stand-alone support vector machines algorithm achieved a macro accuracy of 88.06% and micro accuracy of 96.59% while the random forest algorithm achieved a macro accuracy of 87.33% and micro accuracy of 96.19%.

Keywords: Mitochondrial DNA, Ensemble Learning, Classification Algorithms, Support Vector Machines, Random Forest, Nearest-Neighbor, the Genographic Project, Database Assembly
Associated Activities (4)

Scientific Presentations

EduHPC 2015, Austin TX
Associated Activities (5)

**Scientific Publications in HPC and Bioinformatics**


EduHPC 2015, Austin TX
**Program Outcomes (1)**

*Participants/Offers Made/Applicants*

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th># Female</th>
<th># Underrepresented Minorities</th>
<th># from Target Schools</th>
<th># CS Major</th>
<th># Bio Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>10/14/38</td>
<td>1/4/9</td>
<td>1/4/7</td>
<td>4/7/18</td>
<td>8/12/31</td>
<td>2/2/7</td>
</tr>
<tr>
<td>2013</td>
<td>9/16/117</td>
<td>2/6/36</td>
<td>1/6/18</td>
<td>6/9/42</td>
<td>7/9/72</td>
<td>2/7/45</td>
</tr>
</tbody>
</table>

EduHPC 2015, Austin TX
Program Outcomes (2)

• Graduate School Fellowships
  • 2012 NSF Graduate Research Fellowship
  • 2012 Alfred P. Sloan Foundation Fellowship
  • 2012 NSF Graduate Research Fellowship (Honorable Mention)
  • 2010 GAANN Fellowship, US Dept. of Education

• Graduate School Attending
  • 80% (12-14), 70% (08-10)

• Student Publications and Presentations
  • Several Journals and Conferences in HPC and Computational Biology
REU Site: Bio-Grid Initiatives for Interdisciplinary Research and Education

- Thanks for your time with us at EduHPC15!