

Bioinformatics Data Integration Practicum

03-513 6 units (for undergraduate students) / 03-713 6 units (for graduate students)

Spring first half mini-course

Instructor: Robert F. Murphy

This course will provide a practical experience in integration of bioinformatics data of diverse types in collaboration with a major industrial or government partner. At the beginning of the semester, students will be presented with a description of the problem and sample data sets. During the semester, students will work as part of independent teams to design, implement and evaluate an appropriate data integration system (with the opportunity for interaction with company developers for advice and feedback). The course grade will be based on an oral presentation of the developed software system and a written report describing its development and evaluation. Selected students will have the opportunity to present their work to the company. Prerequisites: 03-310 or 03-311 or 03-510 and 15-211 (15-415 or 15-451 recommended), or permission of instructor.

History

This course was piloted in Spring 2003 as a special section of the Biological Sciences Independent Study course (03-410) and of the Masters Research course (03-700) in collaboration with GlaxoSmithKline. Six students participated in teams of two and five of those students traveled to Philadelphia at GSK expense to present their project results. GSK staff were impressed by the presentations and they would like to continue the course. The student reaction to the course was also very positive. The course was run as a second-half mini course, but discussions with students and GSK led to the conclusion that it would be better to do it as a first-half mini course so that students would have time to arrange possible internships or jobs for the following summer. The material below is a description of the course last year. We would anticipate making similar arrangements with other companies in the future.

Background

Genetics Knowledge Management (GKM) is a data integration system developed at GlaxoSmithKline (<http://www.gsk.com>) that allows scientists to correlate and analyze data from disparate scientific technologies. Scientists routinely analyze data across technologies, yet the effort required to do so on a high-throughput scale is prohibitive, resulting in incomplete understanding of the results. GKM provides scientists a way to combine data sets from multiple high-throughput data sources in a way that preserves the integrity of the original data, yet allows correlations across the data sets.

The underlying data merging processes take place in large part on the scientist's desktop computer using basic (and slow) algorithms, which works well for small data sets. A proprietary technique was implemented to perform these operations on the server-side for larger data sets, yet was not optimized.

Project Task

The task in this project course is to develop an optimal technique for data integration from multiple sources that preserves the ability to analyze data.

The stepwise process to develop a solution includes the following steps: understand the problem and requirements; set up a test case scenario using a small sample data set to verify data integrity; design a highly optimized solution for delivery of large data sets in minimal time;

implement the design; and evaluate the results. Students are encouraged to bring their skills directly into the development phase, either from biological science or computer science field.

The project must result in a working system to be demonstrated during an oral presentation near the conclusion of the course. Also required is a written report describing one or more of the following: metrics for a variety of tested solutions; a model for interfacing the merge component(s) with an existing web services data delivery engine (GKM); or a model for delivering a scalable version of merge, involving high usage and high-volume data sets running in parallel on a server.

Students with an understanding of genomics technologies (gene expression data, protein-protein interaction data, gene sequence annotation data) may add innovative steps to enhance the resultant data set, and allow deeper analyses by scientists. These may include drawing on web-based biological databases. Students with a strong computer science background may develop techniques for managing large data sets, apply parallel processing, test implementation strategies, and focus on application interface issues. Students with interest and background in both areas may ultimately create a “workflow engine” capable of streaming numerous data sets into one highly enriched data set ready for deeper analysis, yet deliver the data in a minimum of time.

Interaction with GlaxoSmithKline

Students may interact with GlaxoSmithKline bioinformaticists and application developers to develop solutions. At the end of the semester, selected students will present their solutions to the GlaxoSmithKline evaluation panel consisting of, but not limited to, members from the GKM development team and representatives from the Information Technology Development Program.