Poster D-13

Ad Hoc Biomedical Computational Tools: An Unfilled Niche

Authors:
Craig R. Street (*Biomedical Informatics Facility, Department of Pathology and Laboratory Medicine, University of Penn*)
David Wang (*Biomedical Informatics Facility, Department of Pathology and Laboratory Medicine, University of Penn*)
David Birtwell (*Biomedical Informatics Facility, Department of Pathology and Laboratory Medicine, University of Penn*)
Kevin Lux (*Biomedical Informatics Facility, Department of Pathology and Laboratory Medicine, University of Penn*)
Alison W. Loren (*Division of Hematology and Oncology, Department of Medicine, University of Pennsylvania, Philadelphi*)
Eline L. Prak (*Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104*)
David Fenstermacher (*Biomedical Informatics Facility, Department of Pathology and Laboratory Medicine, University of Penn*)

Short Abstract: Biomedical Researchers are frustrated managing their data. Our facility at Penn is devising a unique approach in solving local problems. We are utilizing computational technologies to store, mine and retrieve data from genomic, proteomic, and molecular biology experiments. Our development is open-source and can be adopted by groups elsewhere.

Long Abstract:
Biomedical Researchers working in a variety of disciplines at the University of Pennsylvania (Hematology/Oncology, Transplantations, Immunology, etc.) and with a wide spectrum of data formats (Flow Cytometry, Array Comparative Genomic Hybridization, Single Nucleotide Polymorphisms, Microarrays, etc.) are experiencing frustration with managing their data to allow for higher-order analyses. This frustration is compounded because their data is growing faster than their analytical capabilities. Although the situation at Penn is not unique, we are devising a unique approach to solving their problems. The Biomedical Informatics Facility (BMIF) is assisting biomedical investigators in utilizing existing and emerging computational technologies in order to store, mine and retrieve data from genomic, proteomic, sequencing, flow, and molecular biology experiments that have been integrated with clinical and phenotypic information. By integrating these disparate types of data, the BMIF is developing a cadre of applications that arm researchers with new tools with which they can link genotype to phenotype.

Biomedical informatics tools enable biomedical investigators to utilize vast amounts of research and clinical data. This is achieved by creating unified data models, standardizing data interfaces, developing structured vocabularies, generating new data visualization methods and capturing detailed metadata for the investigator's research project. Data exchange, integration and analysis are the underlying themes for creating effective

computational resources that support and extend investigative biomedical research.

A common architecture is being used for these various applications. Cross cutting concerns like authentication and data security are shared between components for increased robustness and reliability. Applications are written in a thin-client model with data access being remotely captured via the web. This allows the users to enter data from case report forms remotely. Common functionally like written survey data entry is factored into shared applications in order to maximize the efficiency of application development and promote the creation of a stable bug free application base. Open source development is encouraged by using a Java Struts framework at the application level. The backend database is Oracle 9i (an open-source database such as MySQL could be substituted). Patient information (e.g., name, date of birth, etc.) are encrypted in the database and stored on a separate secure server to comply with HIPAA regulations.

The Department of Hematology/Oncology will utilize the Leukemia Patient Registry to capture patient data based on a series of visits. Data for any visit may include patient demography, diagnosis, disease stage, cytogenetic and/or molecular abnormalities, participation in a clinical trial, and treatment. In addition, we developed a novel way of capturing data associated with a treatment (namely the associated drugs and regimens). Each drug entry is composed of four components: Drug Name, Dosage, Frequency, and Duration. Regimens are built from a combination of different drug entries. By decoupling the creation of drugs and associated regimens from patient data, the application can easily be adopted by other research hospitals and users have the flexibility of adding new drugs and creating new regimens autonomously. Treatments includes the Regimen, any associated dosage modifications, the number of cycles, and physicians comments. This allows for a detailed and longitudinal history for each patient.

Flow cytometry can be used to collect data on cell size, shape, and the presence or absence of specific markers. Over time the number of antibody reagents that recognize particular cell markers (which can be analyzed not only on the cell surface but also intracellularly) has increased. Furthermore, the number of parameters that can be analyzed simultaneously for a single cell has increased, resulting in large, multi-dimensional data sets.

Multiparameter flow cytometry is often used to monitor lymphocyte immunophenotypes in clinical trials. Given the complexity of flow cytometry data, it has been difficult to correlate these data with clinical and other laboratory data. To address this problem, members of the Bioinformatics group at the University of Pennsylvania have developed a web-based flow cytometric data entry system and relational database. The application allows the analyst to design each panel and the data fields (e.g., number of T cells that CD4+) that are to be stored. Panel information includes the number of tubes, the acquisition gate, and each fluorochrome-antibody conjugate. The fields for analyzed data are open to the investigator's needs and are not restricted in name or quantity. The analyst setting up the fields, however, must include the tube number from which the data is derived and whether the measurement is percent, mean fluorescent intensity, or total number of cells. We are also in the process of developing a means for the analyst to generate PDF reports that can be reviewed and sent to the Principal Investigator of the study. The report generating method will contain flags, which will automatically generate text messages for values out of range, general alerts, etc.

Other projects currently under development or in the requirements gathering stage include a breast cancer clinical trial that will require the integration of clinical data with molecular and sequencing data, a study measuring various clinical data as an early indicator of sepsis, and integration of data from various external sources for an organ transplant group. Our development is open-source, open access and the paradigm used at Penn could be adopted by biomedical informatics groups at other research universities.