COMPUTATIONAL CHALLENGES OF MASS PHENOTYPING

LAWRENCE HUNTER

Computational Bioscience Program
University of Colorado School of Medicine
Aurora, CO 80045 USA
Email: Larry.Hunter@UCDenver.edu

One of the primary challenges in making sense the dramatic increase in human genotype data is finding suitable phenotype information for correlational analyses. While the price of genotyping has fallen dramatically and promises to continue to decrease, the cost of generating the phenotypes necessary to take advantage of this data has held steady or even increased. Until recently, human phenotype data was primarily derived from assays or measurements made in clinical or research laboratories. However, laboratory phenotyping is expensive and low-throughput. Recently, a variety of promising alternatives have arisen that can provide important new information at greatly reduced costs. However, the nature, extent and complexity of the data produced involve significant new computational challenges.

This workshop will begin with an introduction to some of the new modalities, which include: automated abstraction of information from electronic medical records, data streams from medical instruments (e.g. in an intensive care unit) and implanted devices (e.g. cardiac assist devices), data produces by patient social networks, and data from a new generation of inexpensive wearable sensors measuring everything from physical activity to blood glucose.

Most of these new sources of phenotypic data are secondary to some other purpose. Patient records are generated to support clinical care and payment for medical services. Patient social networking sites support patients emotionally and provide peer counseling. Implantable medical devices produce data streams that meet manufacturers' or caregiver requirements. Wearable sensors satisfy personal curiosity or monitor disease progress. Each of these also produces valuable information for genotype correlations.

We will focus on defining the computational challenges arise in the collection, storage, processing, analysis and, especially, in the useful integration of these many new sources of phenotype data into derivatives that facilitate scientifically or medically valuable correlations with genotype. Computational challenges arise due to the diverse nature of the types of data that characterize human phenotypes, the fact that most phenotyping is a secondary use of data produced for other purposes, and the need to integrate, abstract and summarize data in ways that are likely to show correlations with genotype. There are also bioethical challenges in data sharing, anonymization, openness / privacy, consent, and related topics where computational methods might help address other concerns.

The new sources of phenotypic information produce data at radically different time scales and granularities. Modern medical instruments can produce data streams at 50Hz or greater sampling frequencies for days at a time. Patient social networking users typically update their entries every few days, but can be maintained for many years. Effectively integrating information that is produced at such different resolutions and durations is a difficult task. Sensor fusion approaches from other domains may be relevant, although some problems (and some solutions) may be specific to the biomedical domain. Similarly, signal processing approaches that summarize high frequency data into scalar or categorical values may prove of value in this application.

Many of the new sources of phenotypic information produce unstructured or semi-structured data, such as physician notes in electronic medical records, or postings to patient social networking sites. Biomedical natural language processing (NLP) techniques have shown some value in systematizing and normalizing this kind of textual information, but most research in this area has been for clinical decision support, information retrieval, or information extraction. Performance of NLP tools is too often modest in those applications. Are there aspects of using unstructured information to define phenotype that differ from these other applications? Are there differences that can be exploited to improve performance?

Many interesting precedents for the sort of genetic research that these new sources of phenotype data make possible can be found in traditional epidemiology; so can many of the challenges. One particularly pernicious problem in using observational data is the confounding of covariates. For example, many of the patients taking the drug Metformin have elevated blood glucose levels; that's because Metformin is a front-line drug for diabetes, not because taking the drug increases blood sugar. People who wear activity monitors are more active than those that do not, but just giving everyone an activity monitor is unlikely to increase the level of physical activity in the population. Identifying and normalizing for covariates is a critical task in taking advantage of the growth of phenotype data gathered secondary to some other purpose (such as patient care or finding social support). Integration of new data streams with more traditional epidemiological data types (such as demographics or survey results) are also an interesting area for the development of automated methods.

A different class of computational problems arises from the complex personal, social and bioethical concerns around the collection and use of phenotypic data. Are there computational approaches to anonymization, provenance, de-duplication or other problems in making it possible for patients (and normal controls) to share the data they want to with researchers, to protect their rights, to give research participants access to important conclusions drawn around them? Are there developments in electronic consenting, cryptography, or computer security that can facilitate the flow of useful data to researchers while protecting participants?

These topics are relatively new to the computational biomedicine community. The purpose of the workshop is to bring together experts in diverse areas to: identify specific driving problems, define important research topics, and perhaps to share valuable data sets and other research resources. We hope that the outcome of the workshop is a deeper understanding of the challenges, one that will eventually lead to novel computational approaches to addressing these very important problems.