HARNESSING BIG DATA FOR PRECISION MEDICINE: INFRASTRUCTURES AND APPLICATIONS

KUN-HSING YU
Biomedical Informatics Training Program, Stanford University
3165 Porter Dr., Room 2270, Palo Alto, CA 94304
Email: khyu@stanford.edu

STEVEN N. HART
Center for Individualized Medicine, Mayo Clinic
200 First Street SW, Rochester, MN 55905
Email: Hart.Steven@mayo.edu

RACHEL GOLDFEDER
Biomedical Informatics Training Program, Stanford University
870 Quarry Rd, Stanford, CA 94305
Email: rlg2@stanford.edu

QIANGFENG CLIFF ZHANG
School of Life Sciences, Tsinghua University
Medical Science Building, B-1002, Tsinghua University, Beijing, China 100084
Email: zhang.lab@biomed.tsinghua.edu.cn

STEPHEN C. J. PARKER
Computational Medicine and Bioinformatics, University of Michigan
100 Washtenaw Ave, 2049B, Ann Arbor, MI 48109
Email: scjp@umich.edu

MICHAEL SNYDER
Department of Genetics, Stanford University
300 Pasteur Dr., M344 MC 5120, Stanford, CA 94305
Email: mpsnyder@stanford.edu

Precision medicine is a health management approach that accounts for individual differences in genetic backgrounds and environmental exposures. With the recent advancements in high-throughput omics profiling technologies, collections of large study cohorts, and the developments of data mining algorithms, big data in biomedicine is expected to provide novel insights into health and disease states, which can be translated into personalized disease prevention and treatment plans. However, petabytes of biomedical data generated by multiple measurement modalities poses a significant challenge for data analysis, integration, storage, and result interpretation. In addition, patient privacy preservation, coordination between participating medical centers and data analysis working groups, as well as discrepancies in data sharing policies remain important topics of discussion. In this workshop, we invite experts in omics integration, biobank research, and data management to share their perspectives on leveraging big data to enable precision medicine.
Workshop website: http://tinyurl.com/PSB17BigData; HashTag: #PSB17BigData.

1. Introduction

Throughout medicine's history, disease prevention and treatment has been based on the expected outcome of an average patient[1]. Data from patients with the same disease were often pooled together for statistical analysis, and clinical guidelines derived from the aggregated analysis informed health and disease management for billions of patients. Although this approach achieves some success, it ignores important individual differences, which can result in different treatment responses[2].

Precision medicine aims to tailor clinical treatment plans to individual patients, with the goal of delivering the right treatments at the right time to the right patient[3]. Recent advances in omics technologies provide clinicians with more complete patient profiles[4,5]. The decreasing cost of sequencing and associated data storage[6] and the development of effective data analysis methods make it possible to collect and analyze big biomedical data for various human diseases at an unprecedented scale[7]. These advancements can improve the diagnostic accuracy of complex diseases, identify patients who will benefit from targeted therapeutics, and predict diseases before their occurrence[3,8].

Nevertheless, many challenges still remain. Conventional methods for data storage, database management, and computational analysis are insufficient for the petabytes of biomedical data generated every year. In addition, as datasets become larger and more diverse, advanced distributed file storage and computing methods are needed to make the data useful. Furthermore, data-sharing policies and result reproducibility continue to be vigorously debated issues[9-10].

In this workshop, world-renowned experts in personal omics profiling, biobanks, biomedical databases, and medical data analysis will describe recent advancements in these areas and discuss associated challenges and potential solutions.

2. Workshop presentations

This section provides a brief summary for each presentation. The full abstracts could be found at the workshop website http://tinyurl.com/PSB17BigData.

2.1. DeepDive: A Dark Data System

Dr. Christopher Ré, Department of Computer Science, Stanford University, CA, USA

Many pressing questions in science are macroscopic, as they require scientists to integrate information from numerous data sources, often expressed in natural languages or in graphics; these forms of media are fraught with imprecision and ambiguity and so are difficult for machines to understand. Here I describe DeepDive, which is a new type of system designed to cope with these problems. It combines extraction, integration and prediction into one system. For some paleobiology and materials science tasks, DeepDive-based systems have surpassed human volunteers in data quantity and quality (recall and precision). DeepDive is also used by scientists in areas including genomics and drug repurposing, by a number of companies involved in various

forms of search, and by law enforcement in the fight against human trafficking. DeepDive does not allow users to write algorithms; instead, it asks them to write only features. A key technical challenge is scaling up the resulting inference and learning engine, and I will describe our line of work in computing without using traditional synchronization methods including Hogwild! and DimmWitted. DeepDive is open source on github and available from DeepDive.Stanford.Edu.

2.2 Results of the VariantDB Challenge

Dr. Steven Hart, Department of Health Sciences Research, Mayo College of Medicine, MN, USA

The current standard formats for storing genomics data is the VCF and gVCF, but manipulating these large files is an imperfect and impractical long-term solution. Scalability, availability, consistency, are all important drawbacks to the file-based approach. Multiple pieces of metadata are often required to interpret genomic data, but there is no specification for how to tie sample level data (e.g. smoking status, disease status, age of onset, etc.) with variant-level data. The motive of the VariantDB Challenge is to identify a scalable, robust framework for storing, querying and analyzing genomics data in a biologically relevant context. The contextual focus is a central theme in the challenge since it is relatively easy to optimize simple database lookups, but forming queries with multiple predicates becomes a much more complicated task. The VariantDB_Challenge is a 100% open source project, meaning that all code and solutions used must be made publically available via GitHub. In this session, we will present an overview of the challenge and summarize the results from all submitters.

2.3. ADAM: Fast, Scalable Genome Analysis

Mr. Frank Austin Nothaft, Department of Computer Science, UC Berkeley, Berkeley, CA, USA

The detection and analysis of rare genomic events requires integrative analysis across large cohorts with terabytes to petabytes of genomic data. Contemporary genomic analysis tools have not been designed for this scale of data-intensive computing. This talk presents ADAM, an Apache 2 licensed library built on top of the popular Apache Spark distributed computing framework. ADAM is designed to allow genomic analyses to be seamlessly distributed across large clusters, and presents a clean API for writing parallel genomic analysis algorithms. In this talk, we'll look at how we've used ADAM to achieve a 3.5× improvement in end-to-end variant calling latency and a 66% cost improvement over current toolkits, without sacrificing accuracy. We will also talk about using ADAM alongside Apache Hbase to interactively explore large variant datasets.

2.4. Personalized Medicine: Using Omics Profiling and Big Data to Understand and Manage Health and Disease

Dr. Michael Snyder, Department of Genetics, Stanford University School of Medicine, CA, USA

Understanding health and disease requires a detailed analysis of both our DNA and the molecular events that determine human physiology. We performed an integrated Personal Omics Profiling (iPOP) on 70 healthy and prediabetic human subjects over periods of viral infection as well as during controlled weight gain and loss. Our iPOP integrates multiomics information from the host (genomics, epigenomics, transcriptomics, proteomics and metabolomics) and from the gut microbiome. Longitudinal multiomics profiling reveals extensive dynamic biomolecular changes occur during times of perturbation, and the different perturbations have distinct effects on different biomolecules in terms of the levels and duration of changes that occur. Overall, our results demonstrate a global and system-wide level of biochemical and cellular changes occur during environmental exposures.

2.5. Statistical and Dynamical Systems Modeling of Real-Time Adaptive m-Intervention for Pain

Dr. Jingyi Jessica Li, Departments of Statistics and Human Genetics, University of California, Los Angeles, CA, USA

Nearly a quarter of visits to the Emergency Department are for conditions that could have been managed via outpatient treatment; improvements that allow patients to quickly recognize and receive appropriate treatment are crucial. The growing popularity of mobile technology creates new opportunities for real-time adaptive medical intervention, and the simultaneous growth of "big data" sources allows for preparation of personalized recommendations. We present a new mathematical model for the dynamics of subjective pain that consists of a dynamical systems approach using differential equations to forecast future pain levels, as well as a statistical approach tying system parameters to patient data (both personal characteristics and medication response history). We combine this with a new control and optimization strategy to ultimately make optimized, continuously-updated treatment plans balancing competing demands of pain reduction and medication minimization. A workable hybrid model incorporating both mathematical approaches has been developed. Pilot testing of the new mathematical approach suggests that there is significant potential for (1) quantification of current treatment effectiveness for pain management, (2) forecast of pain crisis events, and (3) overall reduction of pain without increased medication use. Further research is needed to demonstrate the effectiveness of the new approach for each of these purposes.

2.6. Integrated Database and Knowledge Base for Genomic Prospective Cohort Study: Lessons Learned from the Tohoku Medical Megabank Project

Dr. Soichi Ogishima, Tohoku Medical Megabank Organization, Tohoku University, Japan

The Tohoku Medical Megabank project is a national project to revitalize medical care and to realize personalized medicine in the disaster area of the Great East Japan Earthquake. In our prospective cohort study, we recruited 150,000 people at Tohoku University, its satellites health clinics, and Iwate Medical University. We collected biospecimen, questionnaire, and physical

measurement during baseline and follow-up investigations. Along with prospective genome-cohort studies, we have developed integrated database and knowledge base, which will be the foundation for realizing personalized medicine and disease prevention.

3. Conclusion

Big data in biomedicine presents a great opportunity to understand health and disease states at an unprecedented level. This workshop will highlight landmark achievements in integrative omics studies, biobank research, and novel data mining methods for large datasets. With the growing number and size of biomedical datasets worldwide, we envision that approaches discussed in this workshop will facilitate the development of precision medicine.

4. Acknowledgments

K.-H. Y. is supported by a Howard Hughes Medical Institute (HHMI) International Student Research Fellowship and a Winston Chen Stanford Graduate Fellowship. R.G. is supported by a National Science Foundation (NSF) Graduate Research Fellowship. M.P. is partially supported by National Institutes of Health grants 1U54DE02378901, 5P50HG00773502, and 5U24CA16003605.

5. References

1. Collins FS. Exceptional opportunities in medical science: a view from the National Institutes of Health. *JAMA*. **313**:131-2 (2015).
2. Shastry BS. Pharmacogenetics and the concept of individualized medicine. *Pharmacogenomics J.* **6**:16-21 (2006).
3. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* **372**:793-5 (2015).
4. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell.* **148**:1293-307 (2012).
5. Yu KH, Snyder M. Omics Profiling in Precision Oncology. *Mol Cell Proteomics.* **15**:2525-36 (2016).
6. Stein LD. The case for cloud computing in genome informatics. *Genome Biol.* **11**:207 (2010).
7. Wichmann HE, Kuhn KA, Waldenberger M, et al. Comprehensive catalog of European biobanks. *Nat Biotechnol.* **29**:795-7 (2011).
8. Ashley EA. The precision medicine initiative: a new national effort. *JAMA.* **313**:2119-20 (2015).
9. Longo DL, Drazen JM. Data Sharing. *N Engl J Med.* **374**:276-7 (2016).
10. Ioannidis JP. Expectations, validity, and reality in omics. *J Clin Epidemiol.* **63**:945-9 (2010).