Abstracts

Abstract Index

Research Highlights

Geotagged Tweets as Predictors of County-Level Health Outcomes
Quynh Nguyen

Fine-Mapping of Obesogenic cis-Regulatory eQTL Variants Using High-Resolution Capture Hi-C
David Pan

Predicting Vasospasm After Subarachnoid Hemorrhage Using High-Frequency Physiological Data
Soojin Park

Creating a Standard Programmatic Interface for Genomic Data With the GA4GH API
Benedict Paten

Understanding Cardiovascular Health and Revealing Pathogenic Insights via Text-Mining Approaches
Peipei Ping

Systems Biology, Meet Evolution and Protein Structure for Characterizing Disease Variant Mechanisms
Jeremy Prokop

Multi-Resolution Analysis of Brain Connectivity: Associations With PET-Based Alzheimer's Pathology
Vikas Singh

KnowEnG: Scalable Knowledge-Guided Analysis of Genomic Data Sets on the Cloud
Saurabh Sinha

Hypothesis Fusion to Improve the Odds of Successful Drug Repurposing
Alexander Tropsha

Integrating Data With Epidemic Simulators to Improve Pandemic Preparedness: Chikungunya in Colombia
Wilbert Van Panhuis

QuBBD: SMART - Spatial-Nonspatial Multidimensional Adaptive Radiotherapy Treatment
David Vock

Extraction and Analysis of Signatures From the Gene Expression Omnibus by the Crowd
Zichen Wang

Data Commons

Common Credit: Evaluating the Scalability of Open Source Applications Across Cloud-Based Providers
Paul Avillach

The MO-LD Project: Enhancing the FAIRness of Yeast and Other Model Organism Data
Michel Dumontier

The smartAPI Initiative: Making Web APIs FAIR
Michel Dumontier

Building The Commons: Interoperable Big Data Publication and Analytics
Ian Foster

BDbags and Minids: Tools for Managing Complex Big Data Sets
Carl Kesselman

FAIR Dataset Landing Pages, Digital Research Objects, and Software Tools for LINCS and BD2K
Amar Koleti

Cloud-Based Drag-and-Drop Scalable RNA Sequencing Pipeline
Alexander Lachmann

The Harmonizome: A Collection of Processed Datasets Gathered to Serve and Mine Knowledge About Genes
Avi Ma'ayan

Reproducibility in Biomedical Sciences
Wladek Minor

Large-Scale, Cloud-Based Analysis of Cancer Data
Brian O'Connor

Automated Deployment of KnowEnG Portal Via Docker Containers in AWS Cloud
Pramod Rizal

Catalyzing Biomedical Research Through the NIH Commons Credits Cloud Computing Paradigm
David Tanenbaum

RNA-seq Pipeline Tutorial With an Example of Reprocessing Data From a Recent Zika Virus Study
Zichen Wang

Deriving Signatures of Pharmacological Action via LINCS Signatures
Lixia Zhang

Big Data for Discovery Science (BDDS): Neuroimaging PheWAS
Lu Zhao

Standards Development

Integrative Representation and Analysis of the LINCS Cell Lines Using the Cell Line Ontology
Caty Chung

Making Phenotypic Data FAIR++ for Disease Diagnosis and Discovery
Melissa Haendel

An Urban Dictionary of Identifier Syntax for the Data Integration Jungle
Julie McMurry

BioSharing - An Informative and Educational Service for Community-Developed Standards
Susanna-Assunta Sansone

Training & Workforce Development

GUI Design and Big Data Visualization of BigDataU Website Development
Sumiko Abe

ERuDIte: The Educational Resource Discovery Index for Data Science Learning
Jose-Luis Ambite

Training Component Activities at the Center for Causal Discovery
Takis Benos

Preparing Underrepresented and First-Generation Students for Careers in Biomedical Big Data Science
Judith Canner

Biomedical Big Data Training for Novices: Initial Experience With a Short-Term Summer School
Brian Chapman

Decaying Relevance of Clinical Data When Predicting Future Decisions
Jonathan Chen

Training and Implementing Genomic Big Data Courses at Primarily Undergraduate Serving Institutions
Jeffrey Chuang

Big Data Research and Education Program in a Primarily Undergraduate Institution (PUI)
Math Cuajungco

Biomedical Research Data Management Open Online Education: Challenges and Lessons Learned
Julie Goldman

Demystifying Biomedical Big Data: A Free OnLine Course
Yuiry Gusev

Data Science Education With MOOCs and Active Learning
Rafael Irizarry

Community Training and Outreach Activities of the BD2K-LINCS DCIC
Sherry Jenkins

Community Research Education and Engagement for Data Science
Patricia Kovatch

Get Real: A Synthetic Dataset Illustrating Clinical and Genetic Covariates
Ted Laderas

Getting Your Hands Dirty With Data
Ted Laderas

Engaging and Training Undergraduates in Big Data Analysis Through Genome Annotation
Wilson Leung

KnowEnG tools for Barrier-Free Learning of Genomic Data Clustering
Mohith Manjunath

Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K) in Puerto Rico
Patricia Ordóñez

Preparing Medical Librarians to Understand and Teach Research Data Management
Kevin Read

MD2K Center of Excellence: Training and Development of a Transdisciplinary mHealth Workforce
Vivek Shetty

The BD2K Training Coordinating Center: A Resource for the Data Science Community
John Van Horn

Data Science Educational Resources for Anyone, Anywhere
Nicole Vasilevsky

University of Washington's R25 Short Course
Daniela Witten

Pandem-Data: Using Big Data in High School
Chuck Wood

Big Data Training for Translational Omics Research
Min Zhang

BioCADDIE & Resource Indexing

Aztec: A Cloud-Based Computational Platform to Integrate Biomedical Resources
Brian Bleakley

The bioCADDIE Data Citation Implementation Pilot
Tim Clark

Reactome: A Curated Knowledge Base of Biomolecular Pathways
Antonio Fabregat

SATORI: A System for Ontology-Guided Visual Exploration of Biomedical Data Repositories
Nils Gehlenborg

A Framework for Metadata Management and Automated Discovery for Heterogeneous Data Integration
Ramkiran Gouripeddi

A Machine Learning Approach for Data Source and Type Identification to Support Metadata Discovery
Ramkiran Gouripeddi

A Scalable Dataset Indexing Infrastructure for the bioCADDIE Data Discovery System
Jeffrey Grethe

Deep Learning-Based Multi-Modal Indexing of Heterogeneous Clinical Data for Patient Cohort Retrieval
Sanda Harabagiu

Omics Discovery Index – Discovering and Linking Public 'Omics' Datasets
Henning Hermjakob

The LINCS Data Portal and FAIR LINCS Dataset Landing Pages
Amar Koleti

Augmenting the the Capabilities for Semantic Search of the Medical Literature
Ani Nenkova

bioCADDIE: Progress and Next Steps
Lucila Ohno-Machado

The DataMed DATS Model, Annotated With Schema.org
Susanna-Assunta Sansone

Reactome: New Services and Widgets to Ease Third-Party Integration
Konstantinos Sidiropoulos

Indexing Clinical Research Datasets Using HL7 FHIR and Schema.org
Harold Solbrig

Datasets2Tools: Enriching DataMed With Canned Analyses
Denis Torre

Aztec I: Building a Technology Platform to Integrate Biomedical Resources
Justin Wood

Development of DataMed, a Data Discovery Index Prototype by bioCADDIE
Hua Xu

Metadata Mapping in bioCADDIE: Challenging Cases
Nansu Zong

Software, Analysis, & Methods Development

Histology-Validated Neural Networks Enable Optical Coherence Tomography Virtual Histology
Vikram Baruah

REproducible by Design, A Docker-Based Tool for Workflow Design, Data Linkage, and Workflow Execution
Tyler Bath

The Georgetown Database of Cancer (G-DOC): A Web-Based Data Sharing Platform for Precision Medicine
Krithika Bhuvaneshwar

Flexible Bootstrapping Approaches Toward the Clustering of Complex Medical Data
Rachael Blair

KnowEnG: Cloud-Based Environment for Scalable Analyses of Genomic Signatures
Charles Blatti

Interactive Web Application for Visualization of Brain Connectivity
David Caldwell

GRcalculator: An Online Tool for Calculating and Mining Drug Response Data
Nicholas Clark

Formal Evidence Networks for Reproducibility in Biomedical Translation
Tim Clark

Augmenting Metadata With Models of Experimental Methods
Scott Colby

Fast, Accurate Causal Search Algorithms From the Center for Causal Discovery (CCD)
Gregory Cooper

Old Medicines, New Uses: Upcycling Drugs Using Social Media and Cheminformatics
Nabarun Dasgupta

A Software Suite for Causal Modeling and Discovery
Jeremy Espino

Clustergrammer: Interactive Visualization and Analysis Tool for High-Dimensional Biological Data
Nicolas Fernandez

Weak Supervision: Biomedical Entity Extraction Without Labeled Data
Jason Fries

Reproducible Exploratory Data Analysis With Vistories
Nils Gehlenborg

TruenoDB: A Network Database System for Managing, Analyzing, and Querying Large Biological Networks
Ananth Grama

Increasing NCBO BioPortal and CEDAR Synergy for BD2K
John Graybeal

Establishing Context: Geospatial Health Context Cube
Timothy Haithcoat

Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification
Sanda Harabagiu

BDDS Tools to Enable Transcriptional Regulatory Network Analysis
Ben Heavner

Biobank to Digibank: High-Frequency Mobile Sensor Data Collection for Long-Lasting Research Utility
Timothy Hnat

The Duke Data Service: Building an Infrastructure for Data and Provenance Microservices
Erich Huang

A Computational Framework for Identifying New Treatment Options in Glioblastoma
Haruka Itakura

SigNetA: Web Tool for Network Analysis of Gene Expression Signatures
Rashid Karim

Building Entity Matching Management Systems for Data Science Problems in Biomedicine
Pradap Konda

Dashboard Visualization and Tool Integration for Enrichr
Maxim Kuleshov

Gene Wiki Knowledgebase and Tool Development for Molecular Signatures of Cardiovascular Phenotypes
Jessica Lee

Pathfinder: Visual Analysis of Paths in Biological Networks
Alexander Lex

GEN3VA: Aggregation and Analysis of Gene Expression Signatures From Related Studies
Avi Ma'ayan

A MACE2K Text Mining Tool to Extract the Impact of Genomic Anomalies on Drug Response
A.S.M. Ashique Mahmood

Faster and Better Metadata Authoring Using CEDAR's Value Recommendations
Marcos Martinez-Romero

New Algorithms for RNA-seq and ChIP-seq Data Compression
Olgica Milenkovic

kBOOM! Intelligent Merging of Different Disease Terminologies
Chris Mungall

Using Crowds to Design Biological Network Visualizations
T.M. Murali

ADAM Enables Distributed Analyses Across Large-Scale Genomic Datasets
Frank Nothaft

GTRAC: Fast Retrieval From Compressed Collections of Genomic Variants
Idoia Ochoa

A Standards-Based Model for Metadata Exchange
Martin O'Connor

Mining Electronic Health Records for Possible Drug Repositioning Opportunities
David Page

Dynamic Control Models for Strategic Interaction
John Pearson

Automatic Discovery and Processing of EEG Cohorts From Clinical Records
Joseph Picone

EEG Event Detection Using Deep Learning
Joseph Picone

Scalable EEG Interpretation Using Deep Learning and Schema Descriptors
Joseph Picone

Integrative LINCS (iLincs): Connecting Diseases, Drugs, and Mechanisms of Actions
Marcin Pilarczyk

NetLINCS: Correlation of Chemical Perturbagen and Signatures to Identify Biological Targets
John Reichard

Compressive Structural Bioinformatics: Large-Scale Analysis and Visualization of the PDB Archive
Peter Rose

Scientific Reproducibility Using the Provenance for Clinical and Healthcare Research Framework
Satya Sahoo

Computational Tools and Resources for LINCS Proteomics Data
Behrouz Shamsaei

SAP – A CEDAR-Based Pipeline for Semantic Annotation of Biomedical Metadata
Ravi Shankar

Using Eye Tracking to Enhance Usability of Big Data in Cancer Precision Medicine
Vishakha Sharma

Big Data Contrast Mining for Genetic Distinctions Between Disease Subtypes
Matt Spencer

Enabling Privacy-Preserving Biomedical Data Analytics in the Cloud and Across Institutions
Haixu Tang

Multitask Deep Neural Net Kinase Activity Profiler
John Turner

Patient Linkage Across Research Datasets in a Patient Information Commons
Griffin Weber

Visualizing Healthcare System Dynamics in Biomedical Big Data
Griffin Weber

KEGGlincs Design and Application: An R Package for Exploring Relationships in Biological Pathways
Shana White

CEDAR: Easing Authoring of Metadata to Make Biomedical Datasets More Findable and Reusable
Debra Willrett

BioThings APIs: Linked High-Performance APIs for Biological Entities
Chunlei Wu

Global Detection of Epistasis
Sihai Zhao

Curate Patient-Centric Multi-Omics Data for Precision Medicine
Jun Zhu

Collaborative Presentations

Aztec and CEDAR: Extraction of Digital Object Metadata From Free Text
Brian Bleakley

Leveraging the CEDAR Workbench for Ontology-Linked Submission of AIRR Data to the NCBI-SRA
Syed Ahmad Chan Bukhari

PREFIX: ACCESSION Compact Identifier Resolution: An EBI/CDL Collaboration
Tim Clark

Cloud-Based Integration of Causal Modeling and Discovery Tools With a Unified Patient Research Database
Jeremy Espino

Annual California BD2K Centers Regional Meetings: Building Connections Across Centers
Ben Heavner

Machine Learning in Textual Data of Cardiovascular Disease via Phrase Mining and Network Embedding
Vincent Kyi

Revisions to the Disease Ontology to Support the Alliance of Genome Resources
Elvira Mitraka

When the World Beats a Path to Your Door: Collaboration in the Era of Big Data
Mark Musen

ELIXIR: A European Distributed Infrastructure for Life-Science Information
Pablo Roman-Garcia

Count Everything: Secure Count Query Framework Across Big Data Centers
Ida Sim

FAIR LINCS Data and Metadata Powered by the CEDAR Framework
Raymond Terryn

Worldwide Big Data Collaborations: Examples From ENIGMA, Spanning 35 Countries
Paul Thompson

Sustainability

The Stewardship Gap
George Alter

The Role of Trustworthy Digital Repositories in Sustainability
David Giaretta

Archiving Interpretations of Variants in ClinVar
Melissa Landrum

Combining Protein and Genome Annotation for Interpretation of Genomic Variants
Peter McGarvey

Interoperability of NURSA, PharmGKB, dkNET, and DataMed
Neil McKenna

Interoperability, Sustainability, and Impact: A UniProt Case Study
Cathy Wu

Research Highlights

Prediction of Putative Causal Variants and Genes at BMD GWAS Loci

Basel Al-Barghouthi, University of Virginia

Osteoporosis is characterized by decreased bone mineral density (BMD), a deterioration of bone microstructure and an increased risk of fracture. Osteoporosis is influenced by genetic variation and, in recent years, genome-wide association studies (GWAS) have revolutionized its genetic analysis. The challenge now is to identify causal genes and variants responsible for these associations. Here, we investigated 63 BMD GWAS loci identified by the Genetic Factors for Osteoporosis (GEFOS) consortium in ~80K individuals. Across the 63 loci, we identified 2491 “proxy” variants in strong (r2>0.8) linkage disequilibrium with the most significant variant at each locus. Most (~98%) proxies were noncoding and located in intergenic (~49%) or intronic (43%) regions. Only one proxy was predicted to impact protein function. These data suggest that variants impacting gene regulation are the primary drivers of BMD associations. To prioritize potential causal variants, we utilized Roadmap Epigenomics data to identify BMD proxies overlapping active transcriptional regulatory elements in 127 tissues/cell-lines. Of the total, 795 proxy variants overlapped regulatory elements in at least one of 127 epigenomes. Since we expect causal variants affecting BMD to do so via modulation of gene expression, we utilized Genotype Tissue Expression Project (GTEx) expression quantitative trait loci (eQTL) data to perform a Bayesian test for the colocalization of eQTL with BMD associations. This analysis identified 33 genes regulated by eQTL that colocalized with BMD associations. To illustrate how these data can be used to inform GWAS, we identified a locus on chromosome 12, containing 74 variants associated with BMD, 18 of which overlapped active regulatory elements in osteoblasts. The eQTL analysis identified two high-priority genes, TMEM263 and AK055712, with colocalizing eQTL. Through the integration of epigenetic and eQTL data, we have generated a refined list of potentially causal variants and genes for a large number of BMD GWAS loci.

Identification of Genotype-Phenotype Associations in Phelan-McDermid Syndrome Using Patient-Sourced Data

Paul Avillach, Harvard Medical School; *Maxime Wack; Claire Hassen-Khodja; Megan O'Boyle; Geraldine Bliss; Liz Horn; Andria Cornell; Cartik Kothari; Catalina Betancur; Isaac Kohane*

Phelan-McDermid Syndrome (PMS) is a syndromic form of autism caused by terminal deletions of the long arm of chromosome 22 affecting at least the SHANK3 gene. It variably associates autism, global developmental delay, delayed speech, neonatal hypotonia, and mildly dysmorphic features. Isolated haploinsufficiency of SHANK3 has been shown to be responsible of a subset of PMS features. The PMS International Registry (PMSIR) compiles clinical data in the form of patient-reported outcomes, as well as patient-sourced genetic test results. Data from the PMSIR have been harmonized and integrated into the BD2K i2b2/tranSMART clinical & genomics data warehouse. We conducted genotype-phenotype analyses using regression models associating the deletion size as a predictor of the different clinical outcomes. 156 patients were included, with deletion sizes ranging from 10.34 kb to 9.057 Mb, with 6 patients presenting small isolated SHANK3 mutations. Increased deletion size is significantly associated with delay in gross motor acquisitions, vesicoureteral reflux, socio-emotional and behavioral development delays, verbal speech, mild dysmorphic features (large fleshy hands, dysplastic toenails/fingernails and sacral dimple), and a spectrum of conditions related to poor muscle tone, suggesting the implication of genes upstream of SHANK3. In this study using data from the PMSIR, we demonstrate the use of entirely patient-sourced registry data consisting of PRO items filled by the parents, and curated genetic test reports to conduct genotype-phenotype analyses. Known results are replicated and novel findings show the ability of registry data to uncover new associations between comorbidities and deleted chromosomal regions in PMS.

Understanding Lung Tissue Heterogeneity in Idiopathic Pulmonary Fibrosis

Panayiotis Benos, University of Pittsburgh; *Akif B. Tosun, University of Pittsburgh School of Medicine; Dimitris V. Manatakis, University of Pittsburgh School of Medicine; Milica Vukmirovic, Yale University; Robert Homer, Yale University; Naftali Kaminski, Yale University; Chakra S. Chennubhotla, University of Pittsburgh School of Medicine*

Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive lung disease. IPF is consequence of fibrosis (irregular wound healing) and the microscopic appearance of fibrosis is heterogeneous. Identifying the causal associations between gene expression and histological structures will not only help understand molecular disease mechanisms involved, but it will also provide insights into potential therapeutic targets. The "lung DBP" (Driving Biomedical Problem), which is part of the Center for Causal Discovery (CCD), aims to study the causal genotype-phenotype mechanisms intrinsic to IPF by integrating and co-analyzing clinical variables reflecting disease progression, histopathological patterns in whole-slide H&E stained tissue images, and RNA-seq data collected from the same tissue.

Methods

Lung tissue from IPF patients and controls has been extracted and three slides were cut sequentially. The top and the bottom slides were stained with H&E and scanned, while the middle slide was used to collect RNA-seq data. We developed new computational pathology methods (i) to characterize fibrosis and other salient phenotypic features in IPF tissues, (ii) to quantify disease heterogeneity and (iii) to classify IPF samples from controls. We used a Mixed Graphical Model (MGM) causal algorithm to integrate these histopathological image features with the gene expression data from the same tissue and other clinical variables.

Results

In this poster, we present new computational pathology algorithms to characterize the heterogeneity in the microscopic appearance of fibrosis in IPF whole-slide H&E stained tissue images. Furthermore, we present the genes and gene networks that are causally related to these features and the clinical variables that are indicative of IPF severity. These results constitute a first step in our understanding of the dynamic changes that potentially occur in a progressive fibrotic lung disease.

A Bayesian Causal Inference Method for Identifying Cancer Drivers of Individual Tumors

Chunhui Cai, University of Pittsburgh

Identifying causative somatic genome alterations (SGAs) driving the development of an individual tumor could both provide insight into disease mechanisms and enable personalized modeling for cancer precision medicine. Although methods exist for identifying driver SGAs at the cohort level, few focus on the drivers of individual tumors. Here, we present a Tumor-specific Driver Identification (TDI) method that infers causal relationships between SGAs and molecular phenotypes (e.g., transcriptomic, proteomic, or metabolomics changes) within a specific tumor. We applied the TDI algorithm to 4,468 tumors across 16 cancer types from The Cancer Genome Atlas (TCGA) and identified those SGAs that causally regulate the differentially expressed genes (DEGs) within each tumor. TDI identified 490 SGAs that had a significant functional impact. The TDI list includes most (86%) of the known drivers published by the TCGA network as well as novel candidate drivers. Our computational evaluation of these SGAs and DEGs support that the causal relationships inferred by TDI are statistically robust, and preliminary experimental results support the predictions by TDI.

Finding the Signal in the Noise: Social Media Early Hospital Notification of Mass Casualty Events

Rachael A. Callcut, University of California, San Francisco; *Sara Moore, University of California, Berkeley; Glenn Wakam, University of California, San Francisco; Alan E Hubbard, University of California, Berkeley; Mitchell J. Cohen, University of Colorado*

Introduction

Delayed notification hinders timely hospital based activations in large scale multiple casualty events. Social We hypothesized that Twitter real-time data would produce a unique and reproducible signal within minutes of multiple casualty events and we investigated the timing of the signal compared with other hospital disaster notification mechanisms.

Methods

Using disaster specific search terms, all relevant tweets from the event to 7 days post-event were analyzed for 5 recent US based multiple casualty events (Boston Bombing [BB], SF Plane Crash [SF], Napa Earthquake [NE], Sandy Hook [SH], and Marysville Shooting [MV]). Quantitative and qualitative analysis of tweet utilization were compared across events.

Results

Over 3.8 million tweets were analyzed (SH 1.8 m, BB 1.1m, SF 430k, MV 250k, NE 205k). Original tweets were 45%, retweets 55%. Peak tweets per min ranged from 209-3326. The mean followers per tweeter ranged from 3382-9992 across events. Retweets were tweeted a mean of 82-564 times per event. Tweets occurred very rapidly for all events (<2 mins) and represented 1% of the total event specific tweets in a median of 13 minutes of the first 911 calls. A 200 tweets/min threshold was reached fastest with NE (2 min), BB (7 min), and SF (18 mins). If this threshold was utilized as a signaling mechanism to place local hospitals on standby for possible large scale events, in all case studies, this signal would have preceded patient arrival. Remarkably, the tweet graphic signatures were consistent across disasters [Figure] except SH which had similar signature but, delayed signal initiation.

Conclusions

This first of its kind evaluation of social media data has demonstrated that this mechanism is a powerful, predictable, and potentially important resource for optimizing disaster response. Further investigated is warranted to assess the utility of prospective signally thresholds for hospital based activation.

A Need for Better Data Sharing Policies: A Review of Data Sharing Policies in Biomedical Journals

Robin Champieux, Oregon Health & Science University; *Jessica Minnier, Oregon Health & Science University; Melissa Haendel, Oregon Health & Science University; Nicole Vasilevsky, Oregon Health & Science University*

There is agreement in the biomedical research community that data sharing is key to making science more transparent, reproducible, and reusable. Publishers could play an important role in facilitating data sharing; however, many journals have not yet implemented data sharing policies and the requirements vary across journals. To assess the pervasiveness and quality of data sharing policies, we reviewed the author instructions of 318 biomedical journals. The policies were coded via a rubric indicating if data sharing was required, recommended, or not addressed at all. The data sharing method and licensing recommendations were examined, as well any mention of reproducibility. The data was analyzed for patterns relating to publishing volume, Journal Impact Factor, and publishing models.

11.9% of the journals stated that data sharing was required as a condition of publication. 9.1% of the journals required data sharing, but did not make clear it would affect publication decisions. 23.3% of the journals only encouraged authors to share data. There was no mention of data sharing in 31.8% of the journals. Impact Factors were higher for journals with data sharing policies. Open access journals were not more likely to require data sharing.

Our study showed only a minority of biomedical journals require data sharing, and a significant association between higher Impact Factors and journals with a data sharing requirement. We found that most data sharing policies did not provide specific guidance on the practices that ensure data is maximally available and reusable. As a continuation of this work, we plan to build a public database of journal data sharing policies, and convene a community of stakeholders to further work on recommendations for strengthening and communicating journal data sharing policies.

Analysis of RNA Editing in Cancer Epithelial-to-Mesenchymal Transition

Tracey Chan, University of California, Los Angeles

RNA editing is an important cellular mechanism that alters sequence information in transcripts and potentially influences protein sequences, alternative splicing, translation, and RNA stability. Catalyzed by ADAR enzymes, the most common editing type is deamination of adenosine to inosine (A-to-I), which translational machinery recognizes as guanosine. RNA editing has been implicated in cancer progression and treatment response, through global and single-gene studies. Higher editing levels were detected in many cancers, and certain editing levels were correlated with tumor subtype, stage, patient survival, and drug sensitivity. Furthermore, ADAR knockdown reduced proliferation and promoted apoptosis in breast cancer cells. These studies support the promising direction of developing RNA editing-related cancer diagnostics and treatments, and highlight the need for better understanding of its regulatory mechanisms, functions, and clinical relevance. Here, we focus on RNA editing in Epithelial-to-Mesenchymal Transition (EMT) of cancer, considering that 90% of cancer deaths are due to metastases, and EMT is a major process contributing to metastasis. Required for embryogenesis and wound healing, EMT is reversible re-programming of tightly connected epithelial cells into motile mesenchymal cells. Following EMT in cancer, invasive mesenchymal cells may enter blood vessels and travel to new metastatic sites. Transcription factors, microRNAs, alternative splicing, and epigenetic modifications are well-established regulators of EMT. However, RNA editing in EMT has not been previously studied, to our knowledge. We applied our previously developed methods to detect high-confidence editing sites from RNA-seq datasets of three EMT-induced cell-lines: ZEB1-overexpressing lung cancer cells, TGFB-treated breast cancer cells, and TGFB-treated normal mammary cells. After identifying RNA-DNA differences and filtering out likely artifacts, we identified over 3800 A-to-I editing sites with significant differential editing between mesenchymal and epithelial states. I will present genomic and functional features of these sites and related genes, and demonstrate the relevance of altered RNA editing to EMT and cancer.

Cataloguing and Curating BRCA1/2 Genetic Variation

***Melissa Cline, University of California, Santa Cruz Genomics Institute**; Brian Craft; Zachary Fischmann; Mary Goldman; Charles Markello; Joseph Thomas; Can Zhang; David Haussler; Rachel Liao; Gunnar Rätsch; Benedict Paten*

An estimated 55 to 65% of women who inherit a pathogenic BRCA variant will develop breast cancer by the age of 70, versus 12% of all women. Further, while approximately 1.3% of women will develop ovarian cancer by age 70, this risk rises to 39% and 17% for women with pathogenic BRCA1 and BRCA2 variants respectively. This has led many women to undergo BRCA testing to understand and manage their risk. Unfortunately, many women will be informed that they carry a Variant of Unknown Significance (VUS). Research on BRCA variation was long hindered by a private gene patent. Gene patenting was struck down by the U.S. Supreme Court in 2012, but in its wake, the data on BRCA variation has remained fragmented with no single source for all available data. Consequently, most doctors and geneticists have been working with incomplete information. To address this need, the Global Alliance for Genomics and Health (GA4GH) launched the BRCA Challenge, a consortium with a goal of cataloguing all public knowledge on BRCA variation. Through this effort, we developed BRCA Exchange (http://www.brcaexchange.org), which aggregates and unifies knowledge from genetic repositories including ClinVar, LOVD, 1000 Genomes, ExAC, BIC, exLOVD, ESP and ENIGMA. BRCA Exchange is currently the largest public repository of BRCA variation, with over 16,700 variants. These data are publicly available to view, download, and query via the GA4GH API. Our collaborators in the ENIGMA consortium are utilizing this data for expert curation of BRCA variants. Since we launched BRCA Exchange one year ago, they have tripled the number of variants that they have curated, with expert reviews of close to 3300 variants. Moving forward, we are working with ENIGMA to identify and aggregate the information needed to curate additional variants, including family history and case-level data.

Big Data Imaging Processing & Analysis (BigDIPA)

Michelle Digman, University of California, Irvine; *Charless Fowlkes, University of California, Irvine*

At UC Irvine we have established a national short course in Big Data Image Processing & Analysis (BigDIPA) intended to increase the number and overall skills of competent research scientists now encountering large, complex image data derived from cutting edge biological/biomedical research approaches. Extraction of knowledge from these imaging sources requires specialized skills and an interdisciplinary mindset. Our course is aimed at providing effective training opportunities in this sector of the "Big Data" science community and leverages a strong interdisciplinary group of researchers at UC Irvine with expertise spanning Systems Biology, advanced microscopy techniques and computational image analysis and visualization.

The weeklong in-person course is designed to offer an intense learning experience delivered in a compact time frame, and opportunities to foster interdisciplinary interactions through small team exercises. Beyond providing an intensive on-site training course, all course materials, tutorial exercises, open source software resources and sample datasets are being developed that will be made freely available through on-line distribution to maximize outreach and encourage additional contributions of curated training resources solicited from the community.

The initial offering of the course covers image acquisition, analysis and visualization of molecular interactions and transport during the cell life cycle, statistical and machine learning techniques for mapping cells and tissues, and computational pipelines for handling massive quantities of image data. A particular concrete area explored in depth by students is the recent development of new computational imaging tools based on high-data rate fluorescence correlation. This new approach is referred to as fluorescence Diffusion Tensor Imaging (fDTI), inspired by the Diffusion Tensor Imaging in the MRI field (DTI), and was used by students to reveal the connectivity of anisotropic protein diffusion routes in real-time in living cells.

Predicting Adverse Cardiovascular Events for Tyrosine Kinase Inhibitors From Molecular Features

Anders Dohlman, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Shadia Zaman, Office of Clinical Pharmacology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration; Darrell Abernethy, Office of Clinical Pharmacology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration; Ravi Iyengar, LINCS DTOXS Center, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Cardiotoxicity is a major concern for the FDA when examining new drug applications (NDAs) for the family of anti-cancer tyrosine kinase inhibitors (TKIs). Despite their therapeutic effect, many TKIs are associated with adverse cardiovascular events (ACEs), including hypertension, decreased ejection fraction, and cardiac failure. Some TKIs are associated with these side effects, while others are not, and little is known about the molecular mechanisms underlying these discrepancies. By integrating different types of molecular data collected by the LINCS consortium, and by others, on the response of human cells to treatment with TKIs, we developed machine learning classifiers that can reliably predict the likelihood that a newly developed TKI will induce cardio-toxicity. We use drug-kinase binding data from an enzyme-liked immunosorbant assay (ELISA), L1000 transcriptional profiling of cancer cell lines, and RNA-seq transcriptional profiling of cardiomyocytes, to identify kinase targets and transcriptional signatures associated with cardiovascular toxicity, thereby unraveling the mechanisms underlying the toxic effects of certain TKIs. The biomarkers we discovered can help determine if newly developed TKIs are likely to induce cardiac complications, suggest mechanisms for such toxicities, to guide further research and assist in regulatory vetting before such adverse events appear in patients.

Consumer Wearable Devices for Health Surveillance and Disease Monitoring

Jessilyn Dunn, Mobilize Center and Department of Genetics, Stanford University; *Xiao Li, Department of Genetics, Stanford University; Denis Salins, Department of Genetics, Stanford University; Michael Snyder, Department of Genetics, Stanford University*

Consumer wearable devices provide an unprecedented opportunity for continuous physiologic monitoring of individuals. With the shift in focus from treatment to prevention, utilizing wearable devices data for health surveillance is an attractive model for disease monitoring and prevention. We collected continuous monitoring data from 43 individuals who wore an “mHealth” smartwatch for an average of 11 months. The device measures heart rate, skin temperature, galvanic skin response, and accelerometry (86,400 measurements for each parameter per day). We verified that the smartwatch reproduced known sleep- and exercise-related heart rate and skin temperature responses. To determine whether physiological monitoring using consumer wearable devices can provide clinically relevant insights, we designed an analytic framework to extract personalized outlier periods of abnormal heart rate and skin temperature from the longitudinal data. We are currently using this framework to analyze four individuals with eight total periods of illness during the device-monitoring period. In our pilot study we performed an in-depth analysis of one participant in particular who had very detailed electronic health records (73 days with extensive clinical tests) including a clinically diagnosed illness that occurred during the 603-day device- monitoring period. Integrating mHealth data with the clinical records, we discovered that the most dramatic outlying event, a prolonged period of abnormally high heart rate and skin temperature, corresponded to dates when the subject presented with Lyme Disease. Changes in heart rate during the onset of Lyme disease may indicate Lyme carditis, a serious condition that must be treated rapidly. Our findings demonstrate that wearable devices can provide an opportunity for timely health event alerts. In conclusion, we found that mHealth measurements can be systematically obtained using portable devices to monitor health-related physiology and that the data can be used for real-time assessment of health states outside of the clinic.

Knowledge-Guided Prioritization of Genes Determinant of Drug Resistance Using ProGENI

Amin Emad, University of Illinois at Urbana-Champaign; *Junmei Cairns, Center for Individualized Medicine, Mayo Clinic; Krishna R. Kalari, Center for Individualized Medicine, Mayo Clinic; Liewei Wang, Center for Individualized Medicine, Mayo Clinic; Saurabh Sinha, Department of Computer Science and Carl R. Woese Institute of Genomic Biology, University of Illinois*

Identification of genes whose basal mRNA expression can predict the sensitivity/resistance of tumor cells to cytotoxic treatments can play an important role in individualized cancer medicine. A pretreatment screening of the expression of genes in the tumor tissue can suggest the best course of chemotherapy or can suggest a combination of drugs to overcome chemoresistance. In this study, we developed a computational method called Prioritization of Genes Enhanced with Network Information (ProGENI), to identify such genes by leveraging both the basal gene expressions and prior knowledge in the form of an experimentally verified network of protein-protein and genetic interactions. This method is based on identifying a small set of genes that a combination of their expression and the activity level of the network module surrounding them shows high correlation with drug response, followed by ranking of the genes based on their relevance to this set using random walk techniques.

Our analysis on a dataset comprised of approximately 300 lymphoblastoid cell lines for 24 cytotoxic treatments revealed a significant improvement in predicting drug sensitivity using ProGENI compared to other methods that do not consider network information. A significant improvement was also observed on another dataset from the Genomics of Drug Sensitivity in Cancer (GDSC) database, containing approximately 600 cell lines from 13 tissue types for 139 drugs. In addition, we used literature evidence in addition to siRNA knockdown experiments to confirm the effect of highly ranked genes on the sensitivity of three drugs: cisplatin, docetaxel and doxorubicin. Our results confirmed the role of more than 73% of the genes (33 out of 45) identified using ProGENI in the sensitivity of cell lines to these three drugs. These results suggest ProGENI to be a powerful computational technique in identifying genes that play a key role in determining the drug response.

Toward a Causome of the Brain

Clark Glymour, Carnegie Mellon University; *Ruben Sanchez-Romero, Carnegie Mellon University; Madelyn Glymour, Carnegie Mellon University; J.D. Ramsey, Carnegie Mellon University; Biwei Huang, Carnegie Mellon University; Kun Zhang, Carnegie Mellon University*

The NIH brain initiative aims, among other things, to uncover the causal connections between brain regions that produce both normal and anomalous cognitive and behavioral functioning. Connectivity studies of the human brain using fMRI and other imaging data have almost always taken one of two forms: correlations of signals at the smallest spatial resolution possible ("voxels") or correlations or partial correlations of average values of spatially clustered voxel signals ("Regions of Interest"). Studies of the first kind claim to estimate the "connectome" of the brain. But it is well known that quite aside from leaving directions of influence unspecified, correlation and partial correlation studies may misidentify causal processes. Correlations produce false transitive connections, and partial correlations produce false connections between two voxels that are causes of a third. While statistical methods, such as the PC and GES (Greedy Equivalence Search) algorithms, can find causal connections and have existed for 25 years, they have been insufficiently fast and accurate for whole brain or whole cortex voxel analyses.

Using a recent dramatic speed-up of the GES algorithm that is called FGS, we have produced an estimate of (undirected) causal connections for the entire cortex for each of 60 resting state scans of the same individual (R. Poldrack). Our current work is focused on identifying both direct and feedback (cyclic) relations using the undirected graphical result. Methods for that purpose exist for small models, but need to be speeded up and improved in accuracy.

IRRMC - A Public Database of Macromolecular Diffraction Experiments

Marek Grabowski, University of Virginia*; Dave Cooper; Marcin Cymborowski; Piotr Sroka; Heping Zheng; Wladek Minor*

The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) has grown into a substantial public repository of primary data from protein diffraction experiments. Currently, the database of the resource contains data from 3070 macromolecular diffraction experiments (5983 datasets), accounting for around 3% of all deposits in the Protein Data Bank (PDB), along with the corresponding partially curated metadata. The resource utilizes a distributed storage architecture accommodating a confederation of individual storage servers, which provides both scalability and sustainability. IRRMC is available through its web portal at http://www.proteindiffraction.org, which can be searched using various criteria. The resource is open to submission of data from the community, including datasets that failed to yield X-ray structures to facilitate collaborative efforts to improve protein structure determination methods. Tools for "wrangling" the data and metadata, including an automated reprocessing pipeline are under development. These tools have already identified situations where a modification of the data collection protocol could have significantly increased the quality of data as well as unearthed a number of diffraction datasets that may benefit from reprocessing. The resource provides a way to ensure availability of “orphan” data left behind by for various reasons by principal investigators and/or extinct large structural biology projects.

Predicting Phenotypes of Osteoarthritis Progression

***Eni Halilaj, Stanford University**; Ya Le, Stanford University; Jennifer L. Hicks, Stanford University; Trevor J. Hastie, Stanford University; Scott L. Delp, Stanford University*

Osteoarthritis (OA) is one of the leading causes of disability in older adults. The uncharacterized heterogeneity of this disease remains a confounder in the design of case-control studies and has significantly hindered the discovery of new disease-modifying treatments. The goal of this study was to characterize distinct phenotypes of osteoarthritis (OA) progression and build a model to predict future progression phenotype based on data collected in one visit. To model structural and symptomatic OA progression, we used eight-year data from the Osteoarthritis Initiative—specifically, joint space width measurements from X-rays and pain scores from the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) questionnaire, since they are the most widely used surrogates of structural and symptomatic disease status. We used a mixed-effects mixture model to cluster the functional data and gradient-boosted decision trees (GBT) to build a cross-validated model for predicting progression phenotypes. The analysis was restricted to 1243 subjects who at the enrollment visit were classified as being at high risk of developing OA based on age, BMI, and medical and occupational histories. The majority of subjects (78%) were slow structural progressors or non-progressors, while the rest (22%) were fast progressors, losing nearly 10% of their joint space width per year. Pain progression was highly variable and not associated with joint space narrowing, thus clustering was based primarily on joint space narrowing. Fast structural progression could be predicted with moderate accuracy based on a GBT model of only a few subject characteristics, including weight, sex, smoking, and history of osteoporosis. Cohort selection based on accurate predictive models will make clinical trials more efficacious, significantly reducing the time and cost needed to detect new anabolic or anti-catabolic drug effects. The discovery of more sensitive imaging and biospecimen markers will also improve the performance of these predictive models in the future.

Large-Scale Biological Text Mining: A Data-Driven Approach

***Jiawei Han, University of Illinois at Urbana-Champaign**; Saurabh Sinha, University of Illinois at Urbana-Champaign; Po-Wei Chan, University of Illinois at Urbana-Champaign; Meng Qu, University of Illinois at Urbana-Champaign; Jingbo Shang, University of Illinois at Urbana-Champaign; Yu Shi, University of Illinois at Urbana-Champaign; Xuan Wang, University of Illinois at Urbana-Champaign; Jinfeng Xiao, University of Illinois at Urbana-Champaign; Doris Xin, University of Illinois at Urbana-Champaign*

The majority of massive amount of data in the real world are unstructured or loosely structured text. To unlock the value of these unstructured text data, it is of great importance to uncover real-world entities and their relationships and building semantics-rich, relatively structured information networks.

A lot of text mining tools take human efforts to do manual data curation and extraction of structures from unstructured data. Unfortunately, this can be costly, un-scalable, and error-prone. Recent advances in data-driven, semi-supervised text mining have led to powerful new data mining methods to mine massive collection of biomedical texts from biomedical research literature, including mining PubMed bibliographic data, research papers in PubMed Central, clinical data sets and numerous bio-medical websites and social media related to biomedical sciences, where data related to diseases, drugs, treatments, chemical compounds, proteins, genes, biological pathways can be integrated and extracted from text data and their characteristics and relationships under different conditions can be mined from such datasets as well. Such mined characteristics, structures and relationships can be made available for integration, annotation, and further analysis, including constructing integrated biological information networks.

We have been developing a data-driven approach that mines phrases, types and structures from unstructured text data, including:

- Scalable phrase mining and information extraction methods for biological corpora
- Automated entity extraction and typing in biological text data
- Deep learning and network-based embedding methods for biological entity and relationship discovery
- Integration of biomedical research literature and social media for effective data mining
- Construction of heterogeneous biological networks by text mining and information integration
- Multidimensional information summarization and visualization from text documents

We will report our recent research progress, successful mining methodologies, experimental results, especially the results obtained by collaborating with the researchers at the BD2K-UCLA Heart Center.

BD2K and Global Genomic Data Sharing

David Haussler, University of California, Santa Cruz Genomics Institute; *Benedict Paten, University of California, Santa Cruz; Brian O'Connor, University of California, Santa Cruz; Kevin Osborn, University of California, Santa Cruz; David Steinberg, University of California, Santa Cruz; Melissa Cline, University of California, Santa Cruz; Mary Goldman, University of California, Santa Cruz; Mark Diekhans, University of California, Santa Cruz; Robert Currie, University of California, Santa Cruz*

It is clear to many BD2K centers that genomics has entered a critical phase where research is rapidly being translated into clinical practice, and thus genomics data must be standardized and joined to other clinical data types. At the same time, data is becoming increasingly global as we collectively recognize that it will take truly huge collections to map out the relationships we need to understand between genotype and phenotype.

Addressing the challenge, we at UCSC have forged a partnership with the Global Alliance for Genomics and Health (GA4GH), an organization dedicated to the global sharing of standardized genomics data. The GA4GH has grown to more than 400 institutional members from more than 40 countries. Translational Genomics (UCSC) and other BD2K centers have contributed to GA4GH, often through global demonstration projects like Beacon, BRCA Challenge, and the recently announced Cancer Gene Trust. Additionally, a set of BD2K centers, including UCSC, MD2K, PIC-SURE and bioCADDIE, have adapted (GA4GH) APIs for the secure exchange of genomic, mobile sensor and clinical data ("Count Everything"). These projects collectively demonstrate the power of globally shared data.

To make shared data analysis possible, orthogonal but vital to these projects is the general development of computational environments in the cloud where big data analysis can occur, inspired by the BD2K notion of a Data Commons. Groups within BD2K are harnessing the power of containerization tools like Docker and workflow languages like CWL to lay the groundwork for such an environment that can exist equivalently in many different clouds and data centers, and run modular, standardized workflows on standardized data accessed through widely accepted APIs. The GA4GH Containers and Workflows Task Team, with a BD2K co-leader, is one contributor to this vision, along with the KnowEng BD2K center. Achieving this vision would be transformational for biomedicine.

Modeling Disease Progression From Sparsely Sampled Observations

***Lukasz Kidzinski, Stanford University**; Apoorva Rajagopal, Stanford University; Jennifer Hicks, Stanford University; Michael Schwartz, University of Minnesota; Trevor Hastie, Stanford University; Scott Delp, Stanford University*

In clinical practice, modeling of treatment outcomes is often based on visits that are sparsely scattered in time. Statistical tools, such as sparse functional principal components, can help extract information from these sparse observations, yet they allow only for univariate predictions of a key indicator. Understanding treatment outcomes often requires analyzing the progression of multiple covariates or more complex objects (e.g., functions or images). Moreover, predictions of a complete object of interest can facilitate communication and interpretability of results.

We illustrate the problem using a dataset of 6066 visits of 2898 patients with cerebral palsy who visited Gillette Children's Hospital as part of routine care. For each visit eleven kinematic joint angles (e.g., knee and hip flexion) per leg were observed during gait. We sought to predict the natural progression of these kinematics, as it may improve predictive power of existing models of treatment outcomes, by providing a better baseline for progression of the disease.

To analyze the curves we first project observations into principal component space. These projected observations can be seen as sparse longitudinal multivariate datapoints, sampled from a continuous multivariate process for each patient. We fit a sparse functional principal components model to explain each patient's variability with age. We fit the model separately to each component, leveraging the fact that projected observations have uncorrelated covariates. Prediction in the projected space can then be inverted to the original space of kinematics curves.

The model will allow us to improve existing models of treatment outcome by incorporating the effect of natural progression in the prediction of long-term effects of a therapy. Analysis of entire curves instead of derivative features, together with appropriate visualization techniques using musculoskeletal simulation software such as OpenSim, facilitates qualitative interpretation and communication of results.

Using Twitter to Study Autism

Denise McGinnis, Boston Children's Hospital

The increasing use of social media as a platform of communication and network building offers a unique opportunity to identify and characterize populations of users. Twitter and other social media outlets have been utilized by those with behavioral disorders such as autism spectrum disorder (ASD) and Asperger's syndrome as a way to build friendships, networks and otherwise communicate with a larger cohort of people. The benefits or drawbacks of these types of interactions have yet to be elucidated but provide an opportunity to analyze population characteristics and behaviors. In this analysis, manual curation was used to isolate a cohort of 138 Twitter users who self-identify as having autism. Preliminary results show differences in lexical diversity, length of tweets and topics discussed when compared to a control population of tweets defined as a cohort of Twitter users who do not self-identify as having autism. Ongoing analyses are working to establish an algorithm that can identify potential autistic Twitter users based on the contents of their tweets. This will allow for aggregate level population analyses on social media patterns of users with ASD, times and frequency of use, location by zip code, size of networks and topics of discussion. This could provide insights to health professionals and caregivers on needs and possible interventions as evidenced by the autistic community.

Geotagged Tweets as Predictors of County-Level Health Outcomes

***Quynh Nguyen, University of Utah**; Matt McCullough, University of Utah Department of Geography; Hsien-Wen Meng, University of Utah Department of Health, Kinesiology, and Recreation; Debjyoti Paul, University of Utah School of Computing; Dapeng Li, Michigan State University Center for Systems Integration and Sustainability; Suraj Kath, University of Utah School of Computing; Elaine Nsoesie, University of Washington Department of Global Health; James VanDerslice, University of Utah Department of Family and Preventive Medicine; Ken Smith, University of Utah Department of Family and Consumer Studies, and Huntsman Cancer Institute; Ming Wen, University of Utah Department of Sociology; Feifei Li, University of Utah School of Computing*

Background

Contextual factors can influence health through exposures to health-promoting and risk-inducing exposures. Nonetheless, the scarcity of consistently constructed contextual data limits understanding of contextual effects and geographical comparisons. Also, the environment is more than its physical features; social processes can affect health through the maintenance of norms, stimulation of new interests, and dispersal of knowledge.

Objective

Our aim was to build a national database from geotagged Twitter data with small-area indicators of prevalent sentiment and social modeling of health behaviors. We then examined whether Twitter characteristics predicted health outcomes.

Method

Between April 2015 and March 2016, we collected and spatially mapped 80 million publicly available geotagged tweets. We classified tweet sentiment using a Maximum Entropy classifier. Using a list of 1430 popular foods and 376 popular physical activities, we tracked the frequency of their social media mentions. In linear regression models, we used Twitter-derived indicators to predict health outcomes across 3000 US counties, controlling for county-level demographics and adjusting standard errors for clustering of county values at the state level. All variables were standardized to have a mean of 0 and standard deviation of 1.

Results

Higher percent happy (-0.07 SD), food (-0.14 SD), and physical activity (-0.12SD) tweets were associated with lower premature mortality. Higher prevalence of food tweets (B= -0.18 SD) and healthy food tweets (B= -0.09 SD) were associated with lower county-level obesity. Conversely, higher caloric density of Twitter food mentions (B= +0.08 SD) was related to higher county-level obesity. Higher prevalence of food tweets (B= -0.13 SD) and physical activity tweets (B= -0.12 SD) were related to lower county-level diabetes.

Conclusion

Social media represents a cost-efficient data resource for the construction of neighborhood features that, in turn, may influence community-level health outcomes.

Fine-Mapping of Obesogenic cis-Regulatory eQTL Variants Using High-Resolution Capture Hi-C

David Pan, University of California, Los Angeles (UCLA); *Kristina Garske, Department of Human Genetics, David Geffen School of Medicine at UCLA; Marcus Alvarez, Department of Human Genetics, David Geffen School of Medicine at UCLA; Chelsea K. Raulerson, Department of Genetics, University of North Carolina; Karen L. Mohlke, Department of Genetics, University of North Carolina; Markku Laakso, Department of Medicine, University of Eastern Finland and Kuopio University Hospital; Päivi Pajukanta, Department of Human Genetics, David Geffen School of Medicine at UCLA, Bioinformatics Interdepartmental Program, UCLA, Molecular Biology Institute at UCLA*

In obese adipose tissue, adipocytes continue to enlarge and eventually burst, causing a cellular inflammatory response, which increases tissue heterogeneity. To identify regulatory variants in a tissue-specific manner across the genome, we performed a cis expression quantitative trait locus (eQTL) study using RNA-sequence data on 793 subcutaneous adipose tissue biopsies from the Finnish METabolic Syndrome In Men (METSIM) cohort. However, linkage disequilibrium and the sheer number of eQTLs make it difficult to pinpoint specific variants for functional study. We hypothesize that many regulatory cis-eQTL SNPs fall inside distal enhancers, looping to physically interact with the promoter of the cis-eQTL SNP and enhancer target gene. We sought to fine-map these cis-eQTL SNPs by searching for the subset of cis-eQTL SNPs located in regulatory elements based on promoter capture Hi-C (CHi-C) data from human white adipocytes. We used the HiCUP and CHiCAGO CHi-C pipelines to identify genomic elements significantly interacting with promoters at HindIII fragment-level resolution with an average fragment size of 4 kb. We compared the location of these genomic elements with METSIM adipose cis-eQTL SNPs to highlight regulatory variants at enhancer sites that may be functionally associated with adipose tissue. Our preliminary results suggest that a non-trivial proportion of gene promoters (15.05%) interact with at least one putative enhancer that contains an adipose cis-eQTL SNP, identifying a set of genes for which the cis-eQTL SNP and enhancer target gene are the same (permutation test, p-value<0.0004). Finally, a subset of these genes showed correlation with BMI (Pearson correlation p-value<3.2×10-7) and significant enrichment in a pathways related to cellular adhesion and glyoxylate and dicarboxylate metabolism (adjusted p<0.01). These genes form our future targets of fine-mapping. Our results can help uncover functional variants in regulatory enhancer-promoter looping interactions relevant for transcriptional regulation in heterogeneous human obesogenic adipose tissue.

Predicting Vasospasm After Subarachnoid Hemorrhage Using High-Frequency Physiological Data

Soojin Park, Columbia University; *Murad Megjhani, Columbia University Department of Neurology; Hans-Peter Frey, Columbia University Department of Neurology; Edouard Grave, Facebook AI Research Group (all work done while postdoctoral fellow at Columbia University Department of Biomedical Informatics); Chris Wiggins, Columbia University Department of Applied Physics and Applied Mathematics; Noemie Elhadad, Columbia University Department of Biomedical Informatics*

Objective

To examine whether patterns within time series data are superior to static admission grading scales for predicting delayed cerebral ischemia (DCI) after subarachnoid hemorrhage (SAH). Consecutive admissions for spontaneous SAH were enrolled in an outcomes study at a tertiary care neuroICU. 390 patients met inclusion criteria. Baseline information and grading scales were evaluated including age, gender, Hunt Hess grade, Modified Fisher scale (MFS), and Glasgow coma scale. An unsupervised approach called Random Kitchen Sink (RKS) extracted features from a universal physiological time series dataset (systolic and diastolic blood pressure, heart rate, respiratory rate, and oxygen saturation). Three different classifiers (Partial Least Squares, Support Vector Machines linear, and Support Vector Machines kernel) were trained using subsets of features: (1) MFS (AUC 0.57), (2) baseline information and grading scales (AUC 0.62), (3) RKS-derived physiologic features (AUC 0.74), and (4) combined baseline information, grading scales, and RKS-derived physiologic features with redundant feature reduction (AUC 0.78). For generalizability, analyses were repeated with less restrictive inclusion criteria with total of 488 patients, and results were as good with an AUC of 0.77. Performance is reported as medians on cross-validation with a 12.5% proportional hold-out set.

Conclusions

After spontaneous SAH, patients are monitored for DCI for up to 14 days. Current prediction tools rely on blood thickness and distribution on admission imaging. While advantageously simple to employ, the MFS predicted DCI in our cohort with an AUC of only 0.57. Adding demographics and other grading scales improved prediction accuracy slightly to AUC 0.62. Adding high-frequency features from a universally obtained physiologic dataset further improved prediction accuracy for DCI to AUC 0.78. With an abundance of data and growing ability for data acquisition and online analysis, there is an opportunity to improve an individual's risk assessment.

Creating a Standard Programmatic Interface for Genomic Data With the GA4GH API

Benedict Paten, University of California, Santa Cruz Genomics Institute; *David Haussler, University of California, Santa Cruz Genomics Institute; Kevin Osborn, University of California, Santa Cruz Genomics Institute; David Steinberg, University of California, Santa Cruz Genomics Institute*

The Global Alliance for Genomics and Health (GA4GH) Genomics API allows for the interoperable exchange of genomic information across multiple organizations and on multiple platforms. This is a freely available open standard for interoperability, that uses common web protocols to support serving and sharing of data on DNA sequences and genomic variation. The API is implemented as a webservice to create a data source which may be integrated into visualization software, web-based genomics portals or processed as part of genomic analysis pipelines. It overcomes the barriers of incompatible infrastructure between organizations and institutions to enable DNA data providers and consumers to better share genomic data and work together on a global scale, advancing genome research and clinical application.

The development of the APIs represents the culmination of contributions from over 15 different task teams of the GA4GH. 100 community contributors helped with the development. The codebase has been forked 230 times. The adoption of the API is growing with major efforts coming from UCSC, Ensembl, and Google. Other institutes are using the data definitions to create GA4GH interfaces including Washington University, Microsoft, Cornell, and UC Berkeley. There is also an international federated compatibility test being run by Aridhia Precision Medicine, Scottland.

The integration team at UCSC is building an ecosystem of tools around the APIs beyond the implementation of the open source GA4GH reference implementation server including a suite of test tools to verify compliance with the GA4GH data definitions, mature python client library, interactive API documentation using Swagger, Dockerization of the reference server, and many more utilities to facilitate adoption of this standard.

We acknowledge the technical and other contributions of the following team members: Jerome Kelleher, Mark Diekhans, Brian Walsh, Sean Upchurch, Danny Colligan, Sarah Hunt, Michael Baudis, Melanie Courtot, Chris Mungall, Andrew Jersaitis.

Understanding Cardiovascular Health and Revealing Pathogenic Insights via Text-Mining Approaches

Peipei Ping, HeartBD2K Center at UCLA; *David Liem, HeartBD2K Center at UCLA; Doris Xin, KnowEnG Center; Quan Cao, HeartBD2K Center at UCLA; Vincent Kyi, HeartBD2K Center at UCLA; Leah Briscoe, HeartBD2K Center at UCLA; Karol Watson, HeartBD2K Center at UCLA; Alex Bui, HeartBD2K Center at UCLA; Jiawei Han, KnowEnG Center*

Over the past decades, mounting information on cardiovascular disease (CVD) from natural language in text is rapidly accumulating. Our pilot study demonstrates the feasibility of applying machine-learning and text-mining techniques on textual data in CVD groups to identify novel classifications, to facilitate predictive analytics, and to aid the clinical decision process. In our study, we applied a combination of phrase-mining algorithms and network-embedding techniques to 551,358 publications (dating from 1995 to 2016) in the Pubmed database, as well as the top-250 proteins that are highly relevant to the cause and treatment of CVD. Six CVD groups are studied, including Cerebrovascular Accidents (CVA), Cardiomyopathies and heart failure (CM), Ischemic Heart Diseases (IHD), Arrhythmias, Valve Dysfunction (VD), and Congenital Heart Disease (CHD).

The top 25 most relevant proteins in CM have a similar score pattern to both IHD and CVA, with the majority of proteins in both heart diseases revealing inflammatory function. Whereas, when we considered CVA and IHD as clustered diseases; VD, CHD, and Arrhythmias share little overlap with those top 25 proteins in CM. Furthermore, contractile protein, Titin, has a high relevance in CM and VD compared to the other CVDs. As expected, Troponin-I has a very high score in IHD. Moreover, Platelet-activating factor acetyl hydrolase, which is a mediator of many inflammatory functions, was identified as relevant for all 6 CVD groups.

Taken together, we demonstrate that a combination of phrase-mining algorithms and network-embedding techniques is effective to recognize hidden patterns underlying textual data contained in Pubmed literature. Novel insights are gained from characterizing these relationships among 250 proteins and 6 major CVDs, offering better understanding of CVD. We believe this new data acquisition strategy will be suitable to extract clinical relevant information from the vast amount of unstructured data in the public domain (e.g., Pubmed).

Systems Biology, Meet Evolution and Protein Structure for Characterizing Disease Variant Mechanisms

Jeremy Prokop, HudsonAlpha Institute for Biotechnology; *Howard J. Jacob, HudsonAlpha Institute for Biotechnology*

Our deep Sequence-to-Structure-to-Function pipeline to characterize genes, the proteins they code, and modifications that alter their function has been applied to ~200 genes within the human genome. Our pipeline consists of evolutionary analysis of codon usage averaging 117 species per gene, conservation motif scanning using the evolution, protein structure modeling, mapping of evolutionary biology onto protein structures, and molecular dynamic simulations. Of the proteins analyzed to date, 39% are transcription factors and 50% are those of complete signaling systems, such as systems involved in cardiovascular regulation. For our transcription factors analyzed, we have included ENCODE ChIP-Seq data into our protein modeling to create molecular visualization of evolution-to-DNA sequence specificity. In particular, the transcription factor modeling has been able to highlight DNA binding sensitivity to CpG methylation in protein families such as the MBD containing genes. The ~15 million codons analyzed for ~150,000 human codon positions have highlighted the evolutionary pattern of ATG and TGG codon conservation (as expected) while also suggesting a deep evolutionarily conserved role of CAG and AAG codons in more than ¼ of all genes analyzed to date. With compiled and databased information for these ~200 genes in combination with labs performing whole genome sequencing for various diseases, we have been able to identify multiple disease mechanisms in genes for Chronic Kidney Disease (SHROOM3), cardiovascular (Renin-Angiotensin system), cancer (BAP1, ASXL1-3, TWIST1-2, SOX1-30), and neurological (MTOR, EBF3, MED13), connecting variants into disease pathways and systems.

Multi-Resolution Analysis of Brain Connectivity: Associations With PET-Based Alzheimer's Pathology

Vikas Singh, University of Wisconsin-Madison; *SeongJae Hwang; WonHwa Kim; Nagesh Adluru; Sterling C. Johnson; Barb B. Bendlin*

Background

Identifying individuals at greatest risk for Alzheimer's disease (AD) at preclinical stages is critical for initiating early treatment. While amyloid accumulation is a primary pathological event in AD, loss of connectivity between brain regions is suspected of contributing to cognitive decline. Even though amyloid pathology is a feature of AD, its effect on connectivity has been difficult to measure in the preclinical (asymptomatic) stage of AD.

Methods

Cognitively asymptomatic participants from the WRAP study (N=135) underwent DTI imaging to assess structural connectivity and PiB PET to measure amyloid accumulation. Connectivity strengths were indexed by mean fractional anisotropy (FA) along tracts connecting 162 gray matter regions. We derived wavelet based multi-resolution connectivity signatures (WaCS) for each connection and modeled its relationship with amyloid accumulation (measured by PiB DVR) in 16 regions that accumulate amyloid plaque in AD. Linear modeling on WaCS yields the p-values.

Results

Amyloid burden was associated with extensive connectivity loss. For example, we found that amyloid deposition in left posterior cingulate is associated with connectivity loss between temporal and occipital regions even in this preclinical stage of AD. The p-values for WaCS for 15 (of 16) PiB-PET ROIs show advantages of multi-resolutional WaCS. For 12 (of 15) ROIs, the significance survives the Bonferroni corrected level. We detected 25 statistically significant (10 of 25 with very strong evidence at the Bonferroni corrected level of 0.01) associations between PiB ROIs and connectivity (7 unique edges).

Conclusions

While prior studies have failed to show a close association between amyloid deposition and structural brain changes, especially in preclinical AD, our new algorithm demonstrates the influence of amyloid burden on structural brain connectivity (in almost all regions implicated as important in AD). Our new results significantly enhance detection of amyloid effects on brain connectivity.

KnowEnG: Scalable Knowledge-Guided Analysis of Genomic Data Sets on the Cloud

Saurabh Sinha, University of Illinois at Urbana-Champaign, KnowEnG BD2K Center

The KnowEnG Center is dedicated to the construction of "Knowledge Engine for Genomics" (KnowEnG), a Cloud-based E-science framework for genomics where biomedical scientists will have access to powerful methods of data mining and machine learning to extract important insights out of genomics data. The scientist will go to the KnowEnG portal with their own data sets in the form of spreadsheets and use KnowEnG to analyze those data sets in the light of a massive compendium of community data sets. These data sets, stored in the form of the "Knowledge Network" – a heterogeneous network of genes and their relationships and annotations – will encapsulate prior knowledge that is incorporated into analysis of user data sets. Docker containers will be used to implement complex analytics workflows on the Cloud.

I will describe current progress of the KnowEnG Center, emphasizing the novel algorithms that we have developed and applied to the discovery of mechanisms underlying diverse phenotypes such as drug response and social behavior. In particular, we have developed (1) a technique based on diffusion component analysis that identifies cancer pathways associated with drug response, (2) an approach that uses network-smoothing of gene expression data and random walks with restart on the Knowledge Network to rank cytotoxicity-related genes, and (3) a probabilistic graphical model that integrates genotype, gene expression and transcription factor-DNA binding data with drug response data to identify regulatory mechanisms of drug response variation across individuals. We have also developed random walk-based methods for gene set characterization, as an alternative to existing techniques such gene set enrichment analysis, and used it to glean systems-level insights about aggressive social behavior.

I will present key ideas of these new approaches to knowledge-guided analysis of omic data sets, as well as major features of the Cloud-based cyber infrastructure enabling these analyses.

Hypothesis Fusion to Improve the Odds of Successful Drug Repurposing

Alexander Tropsha, The University of North Carolina at Chapel Hill; *Eugene Muratov, The University of North Carolina at Chapel Hill; Charles Schmitt, The University of North Carolina at Chapel Hill; Weifan Zheng, North Carolina Central University; Nabarun Dasgupta, The University of North Carolina at Chapel Hill and Epidemico*

The prediction of triangular drug-target-disease (DTD) or drug-target-side effect (DTE) relationships has been at the core of modern rational approaches to drug discovery or reprofiling. These approaches include targeted biological screening; chemical genomics profiling; and, to a smaller extent, text mining to identify well-studied drug-target and target-disease pairs to infer novel drug-disease linkages.

We advocate for an approach that fuses hypotheses generated in the aforementioned types of studies to establishing most reliable and experimentally testable DTD (or DTE) hypotheses. In a case study, we have applied this approach to the discovery of novel anti-Alzheimer indications for existing drugs1. We integrated predictions from two independent approaches: (i) QSAR models, which predicted drugs interacting with serotonin 5HT6 receptor, a known Alzheimer target, and (ii) the cmap approach2 that predicted possible anti-Alzheimer drugs based on anti-correlation of drug-induced gene expression profiles and those of the Alzheimer patients. Selective estrogen receptor modulators (SERMS) have been identified and experimentally validated as 5HT6 binders and memory and cognition enhancers. In another recent study3, we have merged concepts from pharmacovigilance, cheminformatics, and pharmacoepidemiology to identify medications likely to cause Stephens-Johnson syndrome (SJS).

The recently funded BD2K grant (1U01CA207160-01) targets the development of the integrated drug repurposing platform integrating approaches from social media mining, cheminformatics, and text mining toward the discovery of novel indications for existing drugs. We expect to identify (possibly, weak) hypotheses concerning unusual effects of medications reported in the social media. Since biological targets for many diseases are often known, we expect to use cheminformatics approaches to evaluate if direct interaction between drugs implicated in social media in connection with a disease, could bind to the known disease target. In parallel, we also plan to mine biomedical literature and, if possible, electronic medical records, to find confirmation of novel DTD relationships in clinical records.

This work was funded by 1U01CA207160-01.

Integrating Data With Epidemic Simulators to Improve Pandemic Preparedness: Chikungunya in Colombia

Wilbert Van Panhuis, University of Pittsburgh; *Guido Camargo, University of Notre Dame Department of Biological Sciences; Fernando de la Hoz, Universidad Nacional de Colombia Department of Public Health; Donald Burke, University of Pittsburgh Graduate School of Public Health; Michael Wagner, University of Pittsburgh School of Medicine*

Many existing datasets that could be used to counter epidemic threats are not used because standardizing and integrating datasets is time consuming and overwhelming. An epidemic is a complex system comprising pathogen and host populations, mosquitoes, and environmental factors. Figuring out the relationships between these factors is a bottleneck that constrains data integration –and the use of Big Data– for pandemic preparedness. Epidemic simulation models, developed by academic research, can be used to solve the data integration problem because they are blueprints of the epidemic system that mathematically define relationships between all system components. Our innovation is to use a new, ontology-based, data model derived from epidemic simulators to represent datasets in a machine-readable format and to integrate these datasets by simulating a real epidemic. As a use-case, we standardized data about the chikungunya epidemics in Latin America and integrated disease and climate data into an agent-based simulation of the 2014-2016 chikungunya (CHIKV) epidemic in Colombia. This model included 45 million agents that represented the population of Colombia distributed across 10 million locations (households, schools, and workplaces) in ~800,000 grid cells. CHIKV and Zika virus emerged recently in Latin America that has already gained decades of experience with the similar dengue virus (DENV). We found that mosquito-control interventions could have prevented 656,209 clinical cases of CHIKV if implemented by all 770 municipalities with vector populations. Information about previous DENV outbreaks could be used to reduce the scale of the intervention by 60% to 301 municipalities while the impact was reduced by only 25%. This information can help countries with limited resources to target their interventions to the most important hotspots of transmission. Our standard data model can catalyze the integration of existing datasets, which are currently unused, through epidemic simulation into cohesive scenarios for better preparedness against epidemic threats.

QuBBD: SMART - Spatial-Nonspatial Multidimensional Adaptive Radiotherapy Treatment

David Vock, University of Minnesota; *Guadalupe Canahuate, University of Iowa; Clifton D. Fuller, MD Anderson Cancer Center; G. Elisabeta Marai, University of Illinois at Chicago*

Treatment of head and neck cancers involves making multiple treatment decisions (e.g., radiotherapy alone, radiotherapy plus chemotherapy, induction chemotherapy, placing feeding tube, etc.) over time. Such treatment decisions must weigh the often competing goals of improving hard endpoints (e.g., 5-year survival) while minimizing side effects (e.g., dry mouth).

Precision medicine seeks to use Big Data collected from cohorts of patients to develop treatment rules to tailor treatment for new patients based on the their clinical, genomic, imaging, and/or demographic characteristics to improve outcomes. This project aims to develop proof-of-concept computational methodology that will support the construction of a sophisticated precision model of patient-specific outcomes for head and neck cancer therapy.

Although methodologies to estimate the optimal treatment strategy are well-established, there are several aspects inherent in determining the optimal treatment for patients with head and neck cancer which preclude the application of standard algorithms and require the development of innovative approaches among this collaboration among experts from four domains of complementary expertise: radiation oncology, medical imaging, dimensionality reduction, and statistical learning. In particular, we present methodological innovation related to 1) ingesting clinical cohort data; 2) extracting spatial features from imaging data; 3) clustering and dimension reduction for multi-dimensional data; and 4) statistically learning optimal treatment regimes incorporating user preference of efficacy endpoints and toxicity. These methodological advances are implemented and integrated together to form a prototype precision-driven model. The methods developed may be used to derive optimal treatment strategies across not only a variety of cancer diagnoses but other chronic conditions that require making multiple decisions that must weigh the tradeoffs between efficacy and toxicity, including mental health disorders, substance abuse diseases, and diabetes.

Extraction and Analysis of Signatures From the Gene Expression Omnibus by the Crowd

Zichen Wang, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Caroline D. Monteiro, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Gene expression data are accumulating exponentially in public repositories. Reanalysis and integration of themed collections from these studies may provide new insights, but requires further human curation. Here we report a crowdsourcing project to annotate and reanalyze a large number of gene expression profiles from Gene Expression Omnibus (GEO). Through a massive open online course on Coursera, over 70 participants from over 25 countries identify and annotate 2,460 single-gene perturbation signatures, 839 disease versus normal signatures, and 906 drug perturbation signatures. All these signatures are unique and are manually validated for quality. Global analysis of these signatures confirms known associations and identifies novel associations between genes, diseases and drugs. The manually curated signatures are used as a training set to develop classifiers for extracting similar signatures from the entire GEO repository. We develop a web portal to serve these signatures for query, download and visualization. The web portal is available at: http://amp.pharm.mssm.edu/CREEDS

Data Commons

Common Credit: Evaluating the Scalability of Open Source Applications Across Cloud-Based Providers

Paul Avillach, Harvard Medical School; *Chirag Patel; Cartik Saravana; Susanne Churchill; Isaac Kohane*

As part of the NIH BD2K Center of Excellence initiative at Harvard, we have developed the PIC-SURE big data platform http://www.pic-sure.org for the integration of all available biomedical data, including clinical, 'omics, and exposome data, for each patient. PIC-SURE utilizes two principal components to make distributed multidimensional and heterogeneous datasets accessible to authorized users. These include: i) the i2b2/tranSMART software platform, and ii) the PIC-SURE RESTful API that allows platform-agnostic and secure access to the integrated datasets to authorized users. To address analytic reproducibility with the data analysis, our prototype is packaged into Docker containers that can be installed by end users and thereby avoids the hassle of configuring host environments and/or downloading and installing all the software dependencies. We have deployed PIC-SURE on an small EC2 instance of the AWS cloud environment. We integrated reference datasets such as the US CDC and Prevention National Health and Nutritional Examination Survey (NHANES) dataset (http://www.cdc.gov/nchs/nhanes/index.htm) and the Broad Institute's Exome Aggregation Consortium (ExAC) dataset (http://exac.broadinstitute.org/). These resources are not currently available on scalable environment to enable simultaneous access and scalable analytic capability to multiple investigators.

In this Common Credits Model pilot, we proposed to scale up our deployment of the PIC-SURE prototype. We are using Docker Data Center to deploy our Docker containers across both available cloud environments, AWS and IBM SoftLayer. The PIC-SURE module will include: a) the i2b2/tranSMART platform, b) the PIC-SURE RESTful API, and c) the NHANES dataset and d) the ExAC dataset. The Common Credits Model pilot will enable us to evaluate the performance and usability of our platform in the aforementioned cloud environments which we anticipate will be extremely useful for the ongoing sustainability of the program and extensions thereof, as well as for the BD2K Community and beyond.

The MO-LD Project: Enhancing the FAIRness of Yeast and Other Model Organism Data

Michel Dumontier, Stanford University; *Maxime Deraspe, Department of Molecular Medicine, Universite Laval; Gail Binkley, Department of Genetics, Stanford University; Kalpana Karra, Department of Genetics, Stanford University; Gos Micklem, Department of Genetics, University of Cambridge; Julie Sullivan, Department of Genetics, University of Cambridge; Michael J. Cherry, Department of Genetics, Stanford University*

Model organisms such as budding yeast provide a common platform to interrogate and understand cellular and physiological processes. Knowledge about model organisms, whether generated during the course of scientific investigations, or extracted from published articles, are integrated and made available by model organism databases (MODs) such as the Saccharomyces Genome Database (SGD). SGD and many MODs use InterMine, a system for integrating, analysing, and republishing biological data from multiple sources that also enables data-driven bioinformatic analyses through a web user interface and programmatic web services. However, the precise invocation of services and subsequent exploration of returned data require substantial expertise on the structure of the underlying database.

Here, we developed a cloud-ready dockerized platform that uses Semantic Web technologies to transform and make available model organism data in a manner that makes it easier to discover, explore, and query. First, we developed a pipeline to extract, transform, and load a Linked Data representation of the InterMine store. Second, we use Docker to package both software and data for local or remote deployment. Third, we built a lightweight dashboard that packages together existing and SPARQL-aware applications to search, browse, explore, and query the InterMine-based data. Our work extends the InterMine platform, and supports new query functionality across InterMine installations and the network of open Linked Data.

The smartAPI Initiative: Making Web APIs FAIR

Michel Dumontier, Stanford University; *Amrapali Zaveri, Stanford Center for Biomedical Informatics Research, Stanford University; Shima Dastgheib, Stanford Center for Biomedical Informatics Research, Stanford University; Trish Whetzhel, The Scripps Research Institute; Andrew Su, The Scripps Research Institute; Chunlei Wu, The Scripps Research Institute*

Biomedical data analysis is increasingly undertaken using cloud-based, web-friendly application programming interfaces (APIs). However, sifting through API repositories to find the right tools presents a number of formidable challenges: users must not only supply the right combination of search terms to find relevant APIs, but must also closely examine the API outputs to determine how they can be connected together. This task is made more difficult because the APIs generally lack the rich metadata needed to precisely describe the service and the data that it operates on. However, authoring good metadata is seen as tedious and unrewarding, unless they can be demonstrated as useful to users.

The aim of the smartAPI interoperability pilot was to explore the use of semantic technologies such as ontologies and Linked Data for the annotation, discovery, linking, and reuse of smart web APIs. Smart web APIs are easier to discover by enabling more search using their rich semantic metadata, and eliminate data silos by providing Linked Data. We developed the smartAPI metadata specification through a community-based effort that involved surveying and comparing API metadata from multiple repositories as well as multiple API metadata specifications. We extended the Swagger Editor to suggest fields from the smartAPI specification and values from the smartAPI repository and a novel tool to profile the output of an API. The smartAPI repository offers a faceted search user interface coupled with an API for storage, retrieval, field-specific suggestion, and global search over smartAPI descriptions. We are currently working in the context of the CFWG for API interoperability to annotate APIs and evaluate our tools in the context of utility and usability.

Building The Commons: Interoperable Big Data Publication and Analytics

***Ian Foster, University of Chicago**; Kyle Chard, University of Chicago; Ben Heavner, Institute for Systems Biology; Carl Kesselman, University of Southern California; Ravi Madduri, University of Chicago; Arthur Toga, University of Southern California*

The BDDS center is developing technologies that enable rapid and reproducible biomedical discovery. Specifically, we are developing tools and services that support the discovery, exchange, identification, analysis, and publication of big biomedical data. We report here on two capabilities that are central to the NIH's vision of the Commons.

The BDDS Data Repository (BDR) is a scalable, web-based data publication system. Operated as a cloud hosted service for the community, BDR is user managed, meaning that any BD2K center (and others) can easily define data collections: user-managed namespaces in which user-specified policies define the storage location, metadata schema, persistent identifier providers, and access permissions that apply to a set of data. BDR allows collection owners to use their own storage (e.g., institution or project) for publishing data. A web interface and workflows then allow external users to publish data in a collection. BDR also enforces immutability and access control, and indexes associated metadata for discovery.

The BD2KBDDS docker hub addresses the need for reproducible and scalable data analytics. We have created Docker containers for over 500 popular Next Generation Sequencing applications, and best practices analysis pipelines in Globus Genomics for analyzing RNASeq datasets using the containers. A profiling service generates computational profiles of various tools, to enable better utilization of available computational resources. We are working to generate unique identifiers for the containers and analytical pipelines published and accessible from a registry.

We will demonstrate the capabilities of these two systems. We will show how users can define data collections and policies, generate persistent identifiers, and describe datasets. We will also demonstrate how users can publish data to a collection analyze data using elastic docker RNASeq analysis pipelines. We will highlight the use of BDBags as an interoperable data format that enables tight integration between these services.

BDbags and Minids: Tools for Managing Complex Big Data Sets

Carl Kesselman, University of Southern California; *Kyle Chard, University of Chicago; Mike D'Arcy, University of Southern California; Eric Deutsch, Institute for Systems Biology; Ian Foster, University of Chicago; Ben Heavner, Institute for Systems Biology; Ravi Madduri, University of Chicago*

Many domains of science are burdened by multiple "Vs" in the big data continuum. They must frequently manage, share, and analyze complex, large data collections—what we call datasets—that comprise many, often large, files of different types. For example, an imaging genetics study may encompass thousands of high-resolution 3D images, terabytes in size; whole genome sequences, each tens of gigabytes in size; and other heterogeneous clinical data.

Due to their size, complexity and diverse methods of production, files in a dataset may be distributed across multiple storage systems: for example, multiple imaging and genomics repositories. As a result, apparently routine data manipulation workflows become rife with mundane complexities as researchers struggle to assemble large, complex datasets; position data for access to analytic tools; document and disseminate large output datasets; and track the inputs and outputs to analyses for purpose of reproducibility.

While sophisticated conventions have been defined for encoding data and associated metadata, and persistent identifier schemes developed that can bind rich sets of metadata to carefully curated data, all impose significant overheads on the researcher and thus are often not used in practice. We address these issues by proposing simple methods and tools for assembling, sharing, and analyzing large and complex datasets that scientists can easily integrate into their daily workflows. These tools combine a simple and robust method for describing data collections (BDbags), data descriptions (Research Objects), and simple persistent identifiers (Minids) to create a powerful ecosystem of tools and services for big data analysis and sharing. We present these tools and show how they can be used to rapidly assemble complex computational methods over large datasets.

FAIR Dataset Landing Pages, Digital Research Objects, and Software Tools for LINCS and BD2K

Amar Koleti, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; *Raymond Terryn, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, BD2K LINCS Data Coordination and Integration Center; Vasileios Stathias, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, BD2K LINCS Data Coordination and Integration Center; Michele Forlin, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, BD2K LINCS Data Coordination and Integration Center; Dušica Vidovic, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, BD2K LINCS Data Coordination and Integration Center; Caty Chung, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Wen Niu, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, BD2K LINCS Data Coordination and Integration Center; Caroline Monteiro, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, BD2K LINCS Data Coordination and Integration Center; Christopher Mader, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Avi Ma'ayan, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, BD2K LINCS Data Coordination and Integration Center; Mario Medvedovic, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, BD2K LINCS Data Coordination and Integration Center; Stephan Schürer, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, BD2K LINCS Data Coordination and Integration Center*

The Library of Integrated Network-Based Cellular Signatures (LINCS, http://lincsproject.org/) program generates a wide variety of cell-based perturbation-response signatures using diverse assay technologies. For example, LINCS includes large-scale transcriptional profiling of genetic and small molecule perturbations, and various proteomics and imaging datasets. The BD2K LINCS Data Coordination and Integration Center (DCIC) has been developing a collection of tools including data standards specifications, data processing pipelines and infrastructure, a metadata registration system, and a diverse suite of end-user software tools to support and implement an end-to-end solution from submitting LINCS datasets by the Data and Signature Generation Centers (DSGCs) to dataset publication via a Data Portal followed by integrated data analytics enabled by easy to use web-based tools. We will give an overview of LINCS tools with an emphasis on our long-term goal of persistent and FAIR (findable, accessible, interoperable, reusable) LINCS resources by connecting signatures, data processing pipelines, analytes, perturbagens, model systems and related concepts, and analysis software tools via uniquely identifiable digital research objects.

All LINCS Datasets are already indexed in bioCADDIE DataMed. In another example of BD2K and LINCS collaboration, we are working with the CEDAR Metadata Center to develop a LINCS Community Metadata Framework for end-to-end metadata management supporting authoring, curation, validation, management, and sharing of LINCS metadata. Shared metadata facilitated via re-usable, modular, and user-friendly CEDAR templates provide the prospect of cross-searchable linkable datasets connecting many different data generation programs.

In addition to building an advanced integrated knowledge environment, our Center supports several internal and external data science research projects and we have an active outreach and training program. Our software and data analytics resources, data science projects, and training programs are available at http://bd2k-lincs.org/.

Cloud-Based Drag-and-Drop Scalable RNA Sequencing Pipeline

Alexander Lachmann, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Troy Goff, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Caroline D. Montiero, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Denis Torre, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Databases such as the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA) host thousands of RNA-sequencing gene expression studies and are rapidly growing. Such repositories are central for improving our understanding of cellular processes at the global scale. The shift from cDNA microarrays to RNA-seq presents significant challenges when it comes to reanalysis of previously published datasets, and the analysis of new datasets. Compared with microarray data, RNA-seq analysis is orders of magnitude more computationally demanding with respect to raw data size, memory requirements, and computational power. To address the need for developing scalable processing of RNA-seq data, we developed an efficient hybrid-cloud solution. The software architecture of our system is based on dockerized images, simplifying the deployment on any server infrastructure that supports Docker. All software dependencies are contained within the Docker images that can be executed scalably, first locally, and then if demand is high, utilize cloud providers such as Amazon Web Services (AWS). The RNA-seq pipeline supports user interaction via an easy to use web interface and through APIs, enabling researchers to analyze their own high throughput gene expression experiments in context of prior studies. With this pipeline we have reprocessed and analyzed over 8,000 samples from publicly available RNA-seq studies to extract gene expression signatures for drug, disease and single gene perturbations. In addition, we have reprocessed almost all mammalian raw RNA-seq samples currently available on GEO. By the use of a Google Chrome browser extension, we plan to provide links for users to download the processed data in Series Matrix Files format to facilitate data reuse. Currently, the project is part of the first cohort of utilizing the Cloud Credits Model for the BD2K Commons.

The Harmonizome: A Collection of Processed Datasets Gathered to Serve and Mine Knowledge About Genes

Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Andrew D. Rouillrd, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Gregory W. Gundersen, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Genomics, epigenomics, transcriptomics, proteomics and metabolomics efforts rapidly generate a plethora of data on the activity and levels of biomolecules within mammalian cells. At the same time, curation projects that organize knowledge from the biomedical literature into online databases are expanding. Hence, there is a wealth of information about genes with an urgent need for data integration to achieve better knowledge extraction and data reuse. For this purpose, we developed the Harmonizome: a collection of processed datasets gathered to serve knowledge about genes from over 70 major online resources. We extracted, abstracted and organized data into ~72 million functional associations between genes/proteins and their attributes. Such attributes could be physical relationships with other biomolecules, expression in cell lines and tissues, genetic associations with knockout mouse or human phenotypes, or changes in expression after drug treatment. We stored these associations in a relational database along with rich metadata for the genes/proteins, their attributes and the original resources. The freely available Harmonizome web portal provides a graphical user interface, a web service and a mobile app for querying, browsing and downloading all of the collected data. To demonstrate the utility of the Harmonizome, we computed and visualized gene-gene and attribute-attribute similarity networks, and through unsupervised clustering, identified many unexpected relationships by combining pairs of datasets such as the association between kinase perturbations and disease signatures. We also applied supervised machine learning methods to predict novel substrates for kinases, endogenous ligands for G-protein coupled receptors, mouse phenotypes for knockout genes, and classified unannotated transmembrane proteins for likelihood of being ion channels. The Harmonizome is a comprehensive resource of knowledge about genes and proteins, and as such, it enables researchers to discover novel relationships between biological entities, as well as form novel data-driven hypotheses for experimental validation. URL: http://amp.pharm.mssm.edu/Harmonizome.

Reproducibility in Biomedical Sciences

Wladek Minor, University of Virginia; *Marek Grabowski, University of Virginia*

Experimental reproducibility is the cornerstone of scientific research, upon which all progress rests. The veracity of scientific publications is crucial because subsequent lines of investigation rely on previous knowledge. Several recent systematic surveys of academic results published in biomedical journals reveal that a large fraction of representative sets of studies in a variety of fields cannot be reproduced in another laboratory. Big Data approach and especially NIH Big Data to Knowledge (BD2K) program is coming to the rescue.

The goal of the presented research is to provide the biomedical community with a strategy to increase the reproducibility of reported results for a wide range of experiments by building a set of “best practices”, culled by extensive data harvesting and curation, combined with experimental verification of the parameters crucial for reproducibility. Experimental verification assisted by the automatic/semi-automatic harvesting of data from laboratory equipment into the already developed sophisticated laboratory information management system (LIMS) will be presented. This data-in, information out paradigm will be discussed.

Large-Scale, Cloud-Based Analysis of Cancer Data

Brian O'Connor, University of California, Santa Cruz; *Benedict Paten, University of California, Santa Cruz; David Haussler, University of California, Santa Cruz*

Creating mobile workflows and tools designed to work across clouds is a challenge for genomics researchers. In this talk I will describe the efforts of the GA4GH Containers and Workflows task team in creating a standard for this approach. I will detail the creation of the Dockstore, our best-practice platform for exchanging workflows and tools using Docker, and will detail its use as a mechanism to package scientific tools and send them to various clouds. This approach was instrumental for projects like the ICGC's PanCancer Analysis of Whole Genomes (PCAWG), which relied on portable workflows to compute across multiple environments. Furthermore, I will describe Toil project, which is a platform for running tools and workflows at scale on AWS and other clouds. The latest Toil recompute of 20,000+ RNASeq samples shows the power of combining Docker-based tools with a highly scalable workflow engine. Finally, I will explore the future of large-scale genomics work using Dockstore, Toil, Consonance, and related GA4GH standards together, and how this complements the vision of the Commons.

Automated Deployment of KnowEnG Portal Via Docker Containers in AWS Cloud

Pramod Rizal, University of Illinois at Urbana-Champaign; *Omar N. Sobh, Umberto Ravaioli*

Activities of the KnowEnG BD2K Center include workflow development of downstream genomic analysis for cloud deployment. Workflows have been assembled with modular Docker containers, to facilitate scalable cloud implementation. The computational scaffold is built on top of the open source HUBzero cyber-infrastructure, providing a well-tested authentication layer. In the spirit of "Build, Ship, & Run Any App, Anywhere," and serve the widest possible audience, it is desirable to create a complete system that can be easily cloned with independent instances on the computational environment of choice. We wish to maintain the design based on the HUBzero cyber-infrastructure and preserve the secure authentication features for separate organizations, but at the same time one cannot expect most users to be able to deploy and maintain HUBzero on their own. Our solution has been to create a containerized version of HUBzero, suitable for automatic installation in support of KnowEnG applications deployment. We have created a Debian based Docker image that any host can pull and install in a few minutes. A github repository is invoked to feed project specific folders and files into the stock Hubzero CMS, to give the exact KnowEnG look and feel to the cloned portal. With this containerized procedure, we have fully automated HUBzero installation. Currently, we host our demonstration HUBzero docker container within an Amazon Web Services (AWS) EC2 instance. The web server is connected to Amazon Relational Database Service (RDS) so our database is fully backed up and secure. It can be used by any instance of fully automated HUBzero docker container installation. Advantages of the approach include: lightweight hosting of HUBzero portal (no dedicated Virtual Machine to be allocated); ability to restart instances of KnowEnG on demand, avoiding charges during prolong idle periods; fast and automated installation of HUBzero applicable to other project.

Catalyzing Biomedical Research Through the NIH Commons Credits Cloud Computing Paradigm

David Tanenbaum, Centers for Medicare & Medicaid Services Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Centers (FFRDC), operated by MITRE Corporation; *Eldred A. Ribeiro, CAMH FFRDC, operated by MITRE Corporation; Lauren Quattrochi, CAMH FFRDC, operated by MITRE Corporation; Wenling E. Chang, CAMH FFRDC, operated by MITRE Corporation; Peter Gutgarts, CAMH FFRDC, operated by MITRE Corporation; Ari Abrams-Kudan, CAMH FFRDC, operated by MITRE Corporation; William Kim, CAMH FFRDC, operated by MITRE Corporation; Lisa Tutterow, CAMH FFRDC, operated by MITRE Corporation; Erin Williams, CAMH FFRDC, operated by MITRE Corporation*

The National Institutes of Health (NIH) invested more than $30 billion in biomedical research during fiscal year 2015, for which Digital Objects such as data, metadata, software, or workflows are among the highest value creations. Effective reuse of Digital Objects permits the greatest scientific and societal return on the investments made by NIH and other funders of biomedical research. The traditional funding paradigm for Digital Objects relies on locally provisioned information technology resources, but increasing use of high-volume data generation and analytic technologies have strained this model. Access to scalable storage and compute to support research is essential for the success of NIH-funded research programs in the face of ever bigger and richer data and informatics.

NIH is instantiating a cloud-based electronic environment (Commons) where researchers can store, share, and make computations on digital data using sharable software, workflows, metadata, and other Digital Objects. The Commons will support the needs of the biomedical research community at reduced cost by leveraging advances in cloud computing and storage. The Commons will be supported, in part, by NIH-provided resources called Commons Credits.

MITRE is conducting a three-year pilot for NIH to evaluate the use of the Commons Credits Model for obtaining cloud services to perform biomedical computational research. Interested research investigators with NIH grants are encouraged to apply during three open Credits Request Cycles, starting approximately November 30, 2016. This work is sponsored by the Office of the NIH Associate Director for Data Science (ADDS), and is part of the Big Data To Knowledge (BD2K) Initiative.

RNA-seq Pipeline Tutorial With an Example of Reprocessing Data From a Recent Zika Virus Study

Zichen Wang, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

RNA-seq analysis is becoming a standard method for global gene expression profiling. However, open and standard pipelines to perform RNA-seq analysis by non-experts remain challenging due to the large size of the raw data files and the hardware requirements for running the alignment step. Here we introduce a reproducible open source RNA-seq pipeline delivered as an IPython notebook and a Docker image. The pipeline uses state-of-the-art tools and can run on various platforms with minimal configuration overhead. The pipeline enables the extraction of knowledge from typical RNA-seq studies by generating interactive principal component analysis (PCA) and hierarchical clustering (HC) plots, performing enrichment analyses against over 90 gene set libraries, and obtaining lists of small molecules that are predicted to either mimic or reverse the observed changes in mRNA expression. We apply the pipeline to a recently published RNA-seq dataset collected from human neuronal progenitors infected with the Zika virus (ZIKV). In addition to confirming the presence of cell cycle genes among the genes that are downregulated by ZIKV, our analysis uncovers significant overlap with upregulated genes that when knocked out in mice induce defects in brain morphology. This result potentially points to the molecular processes associated with the microcephaly phenotype observed in newborns from pregnant mothers infected with the virus. In addition, our analysis predicts small molecules that can either mimic or reverse the expression changes induced by ZIKV. The IPython notebook and Docker image are freely available at: http://nbviewer.jupyter.org/github/maayanlab/Zika-RNAseq-Pipeline/blob/master/Zika.ipynb & https://hub.docker.com/r/maayanlab/zika/

Deriving Signatures of Pharmacological Action via LINCS Signatures

Lixia Zhang, University of Cincinnati; *Wen Niu, Mario Medvedovic*

LINCS Transcriptional Signatures of Drug Effects (LINCSTD) data contains more than 65,000 reaction conditions, 1538 drug information, 72 cell information and so on. Such a big data allows us to identify sets of genes whose expression pattern correlates with specific pharmacological actions. And ultimately, such gene lists can represent sensitive signatures for predicting the chemical mode of action based on transcriptional signatures.

To identify genes associated with specific pharmacological action, we utilized the random-set test for each of the probes of LINCSTD data against the pharmacological categories that separate transcriptional signatures of the LINCSTD data according to its known pharmacological action. From this application, we found out that under which probe, a specific pharmacological category is enriched. By reorganizing the result of the enrichment, we got the p-value of enrichment for each probe under each pharmacological category. Using these p-values to do the random-set test against a specified group of functional gene sets again, we identified sets of genes whose expression pattern correlates with specific pharmacological actions. To identify the association between genes and transcription signatures, we used the random-set test for each of the transcriptional signatures of LINCSTD data against the specified group of gene set. After multiple test adjustment, we found out the transcriptional signatures, the corresponding cell information and drug information that the group of gene set is enriched in. Furthermore, we tested the utility of such gene list signatures in connecting transcriptional signatures of disease with signatures of pharmacological action of the drugs that are currently used for treatment.

Big Data for Discovery Science (BDDS): Neuroimaging PheWAS

Lu Zhao, University of Southern California; *Kristi Clark, Laboratory of Neuro Imaging, Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California; Carl Kesselman, Information Sciences Institute, University of Southern California; Mike D'Arcy, Information Sciences Institute, University of Southern California; Clio Gonzalez-Zaharias, Laboratory of Neuro Imaging, Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California; Ian Foster, Statistics Online Computational Resource, HBBS, University of Michigan, Ann Arbor; Ivo D. Dinov, Statistics Online Computational Resource, HBBS, University of Michigan, Ann Arbor, Michigan Institute for Data Science, HBBS, University of Michigan, Ann Arbor; Farshid Sepehrband, Laboratory of Neuro Imaging, Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California; Arthur W. Toga, Laboratory of Neuro Imaging, Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California*

The recent emergence of neuroimaging genomic databases of large healthy and diseased cohorts enables new approaches for scientific discovery, known as discovery science, to develop and test new genetic hypotheses against large (existing) data. In this work, we developed a new big data discovery framework for neuroimaging based phenome-wide association studies (PheWAS). As the inverse of genome-wide association study (GWAS), PheWAS explores relationships of a single nucleotide polymorphism (SNP) of interest with a wide variety of brain phenotypes using a unified genotype-to-phenotype strategy. This approach provides a broad survey of possible gene-brain associations on whole populations, and enables systematical validation of findings of existing GWASs and single-phenotype studies and discovery of novel associations. Specially, we developed unified big data digital asset management software to manage large-scale, complex and heterogeneous neuroimaging genomic data archives and to produce the data bag for PheWAS. Complex and sophisticated image processing techniques and tools were applied to extract various brain phenotypes from extensive high-resolution neuroimaging data. We also developed a comprehensive analytical tool to implement the statistical analysis for PheWAS across the brain and visualize the obtained PheWAS data mapping on a standard brain model. The PheWAS approach was applied to mine neuroimaging genomic datasets of two independent large normal neurodevelopmental cohorts to investigate impacts of the Catechol-O-methyltransferase (COMT) gene on brain morphology. We robustly replicated the positive associations of the Met158 allele with cortical thickness and the age-related cortical thinning in the prefrontal cortex, and identified novel COMT-brain morphology associations in the parietal and motor cortices. These results systematically describe how COMT influences brain structure and provide a neuroanatomical basis for COMT influences on cognitive functions and motor control in neurodevelopment, and demonstrate the considerable utility of using PheWAS and big data strategies to explore system-level gene-brain associations.

Standards Development

Integrative Representation and Analysis of the LINCS Cell Lines Using the Cell Line Ontology

Caty Chung, University of Miami; *Edison Ong, University of Michigan, Ann Arbor; Jiangan Xie, University of Michigan, Ann Arbor; Zhaohui Ni, University of Michigan, Ann Arbor; Qingping Liu, University of Michigan, Ann Arbor; Yu Lin, University of Miami; Vasileios Stathias, University of Miami; Caty Chung, University of Miami; Stephan Schürer, University of Miami; Yongqun He, University of Michigan, Ann Arbor*

Cell lines are crucial to study molecular signatures and pathways, and are widely used in the NIH Common Fund LINCS project. The Cell Line Ontology (CLO) is a community-based ontology representing and classifying cell lines from different resources. To better serve the LINCS research community, from the LINCS Data Portal and ChEMBL we identified 1,097 LINCS cell lines among which 717 cell lines were associated with 121 cancer types. 352 cell line terms did not exist in CLO. To harmonize LINCS cell line representation and CLO, CLO design patterns were slightly updated to add new information of the LINCS cell lines including different database cross-reference IDs. A new shortcut relation was generated to directly link a cell line to the disease of the patient from whom the cell line was originated. After new LINCS cell lines and related information were added to CLO, a CLO subset/view (LINCS-CLOview) of LINCS cell lines was generated and analyzed to identify scientific insights into these LINCS cell lines. Furthermore, we used the LINCS-CLOview to link and analyze various data generated from different LINCS cell lines. This study provides a first time use case on how CLO can be updated and applied to support cell line research from a specific research community or project initiative.

Making Phenotypic Data FAIR++ for Disease Diagnosis and Discovery

Melissa Haendel, Oregon Health & Science University; Jules Jacobsen, Queen Mary University; James Balhoff, RTI International; Jeremy Nguyen-Xuan, Lawrence Berkeley National Laboratory; Kent Shefchek, Oregon Health & Science University; Dan Keith, Oregon Health & Science University; Harry Hochheiser, University of Pittsburgh; Suzanna E. Lewis, Lawrence Berkeley National Laboratory; Sebastian Köhler, Charité – Universitätsmedizin Berlin; Peter Robinson, The Jackson Laboratory; Julie A. McMurry, Oregon Health & Science University; Tudor Groza, Garvan Institute of Medical Research, Sydney; Christopher J. Mungall, Lawrence Berkeley National Laboratory

While great strides have been made in exchange formats and data models for genomic sequence and variation data (e.g. Variant Call Format; VCF), the same is not true for phenotypic features. The heterogeneity of phenotype descriptions is a reflection of the different purposes for which they are collected (free-text clinical observations, QTLs, and newer high-throughput phenotype measurements) and the different contexts in which they are communicated (publications, databases, health records, registries, clinical trials, and even social media). Biocuration has effectively overcome these challenges in focused studies, but the field still challenged by the lack of broader phenotype standardization, accessibility, persistence, and computability. Consequently, it is extremely difficult to exchange, aggregate, and operate over phenotypic data in the same way that we do sequence data. We have therefore designed an exchange format standard for flexible, extensible, and expressive representation of a broad range of phenotypes in any species. The Phenotype eXchange Format (PXF) does not stop at just phenotypes; it accommodates any other evidence needed to make the very most of these phenotypes (eg. quantitative measurements, environments, and exposures). The goal is to enable PHI-free open exchange of phenotypic data in a way that can be automatically converted from existing registry/clinic data and easily shared outside paywalls for journals during publication of genotype-phenotype studies. In the context of the Global Alliance for Genomics and Health schemas and APIs, PXF will allow consumption of these phenotypic data for computational use by clinical labs for defining gene panels, for diagnostic pipelines, for rare disease patient matchmaking, and for deposition and aggregation in public knowledge bases such as the Monarch Initiative. Open exchange of phenotypic data promotes algorithmic innovation and sharing of our greater understanding of the correlations between genotype, environmental factors, and phenotypic outcomes in more holistic and translational manner. http://phenopackets.org

An Urban Dictionary of Identifier Syntax for the Data Integration Jungle

Julie McMurry, Oregon Health & Science University; *Chris Mungall, Lawrence Berkeley National Laboratory; Melissa Haendel, Oregon Health & Science University; Michel Dumontier, Stanford University*

Identifiers are essential to the flow and integrity of biomedical data. However, many identifier issues impede integrating data at scale. There is no uniform way to represent the same identifier, whether in short form or long (http); nor is there an established mechanism to convert them. For instance, a single NCBI gene identifier (6622) has been represented using 38 distinct short-form representations and 14 distinct http URIs.

The PrefixCommons project aims to facilitate traversal across connected datasets using JSON-LD context files, which can provision mappings between short names or prefixes to globally unique URIs. Within a context file, any given prefix must have a unique URI expansion, but there is no requirement for prefixes to be unique across contexts. As part of this community standards effort, we have developed tools for composition of context files and for testing consistency of different context files. This allows groups to compositionally create context files for any given use case and provides a framework for iteratively approaching a more unified standard. For example, the prefixes utilized by the NIH Commons do not need to be globally unique, however, they must be unique within NIH Commons so that the mapping between these prefixes and the resolving namespaces provide the appropriate URI/URL. Preliminary work in suggests that the number of true collisions (http://tinyurl.com/prefixcollisions) is much smaller than originally hypothesized: of over 2,200 distinct prefixes, only 14 (0.006%) actually refer to two data different providers. A much more common, but harder-to-solve problem is that, for a given data provider, multiple different prefixes have been registered and in different places. The PrefixCommons effort aims to capture these prefix “synonyms” and provide prospective declarative identifier syntax.

BioSharing - An Informative and Educational Service for Community-Developed Standards

Susanna-Assunta Sansone, University of Oxford; *Peter McQuilton; Alejandra Gonzalez-Beltran; Philippe Rocca-Serra; Milo Thurston; Massimiliano Izzo; Allyson Lister; Melanie Adekale; Delphine Dauga and Community Contributors*

In the life/biomedical sciences there are almost a thousand standards and several thousands databases and tools, designed to assist the virtuous data cycle, from collection to annotation, through preservation and publication to subsequent sharing and reuse. As a consumer of these resources, it is often difficult to know which one are the most relevant for your specific domain and needs; as producers, you want to be sure your resource is findable by prospective users.

With its growing and interlinked content, functionalities and endorsements, BioSharing (https://biosharing.org) is the most comprehensive informative and educational resource of community-developed standards. We works with and for the community to map their landscape and define indicators necessary to monitor e.g. their: development, evolution, integration; implementation and use in databases, tools; and adoption in data policies by funders, journals and other organizations.

Whether you are a researcher, standard/database developer, funder, journal editor, librarian or data manager, BioSharing can help you understand which standards are mature and appropriate to your use case. By mapping the relationships between standards and databases/tools that implement them, or the policies that recommend them, BioSharing enables you to make an informed decision as to which standard or database to use or endorse. It also increases the visibility of standards outside their direct domain, reducing the potential for unnecessary reinvention and maximize their (re)use.

Selected as a resource of the ELIXIR interoperability platform, BioSharing also operates as an open working group in the Research Data Alliance and Force11.

This presentation will show: exemplars on how we help consumer and producer of standards; results of a recent user survey, done in collaboration with NIH BD2K and ELIXIR, helping us improving the service; and engagement in and connection with BD2K and ELIXIR activities, such as Bioschemas.org to enhance discoverability of the resources.

Training & Workforce Development

GUI Design and Big Data Visualization of BigDataU Website Development

Sumiko Abe, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; *John Darrell Van Horn; Jeana Kamdar; Crystal Stewart; Avnish Bhattrai; Xiaoxiao Lei; Krithika Ramaswamy; Shobhit Agarwal; Carmen Tan; Caroline O'Driscoll; Lily Fierro; Jose-Luis Ambite; Kristina Lerman; Jonathan Gordon; Gully Burns; Rochelle Tractenberg; Florian Geigl; Priya Jain; John Berardo; Michael Taylor*

The BD2K Training Coordinating Center (TCC) website (BigDataU.org) is an interface between learners of biomedical data science and a rich database educational resource. Big Data U allows the interested learner to easily and quickly find educational resources on a variety of "big data" topics and themes in order to construct personalized training experiences. Using a unique GUI Design, the learner can easily customize and modify multiple educational and training plans which fit their specific needs and interests. Currently, BD2K TCC maintains two essential databases: 1) the Educational Resource Discovery Index (ERuDIte), a growing resource database which presently includes over a thousand educational resources; and 2) a user database for storing learner information, including a profile and saved educational plans. Using machine learning, information retrieval, and natural language processing methods, ERuDIte automatically provides tags and resources to the learners. Grouped tags form the basis for the learner to explore, assemble resources of interest, and to efficiently meet their educational goals in data science. The interaction between each resource, resource and tags, and from tag-to-tag, may be visualized by interactive, graphical 'Knowledge Maps'. Exploring these maps, the learner can easily begin their educational adventure and reach over a thousand ERuDIte educational resources by the click of a mouse. With an increase in use of BigDataU.org and proper user monitoring, learning trends can be recorded, modeled, and will lead to topic suggestions for interested learners. The BD2K TCC is committed to improving the visualization of this knowledge and to providing a novel user-experience through our BigDataU.org website. With your own BD2K training activities, the ERuDIte resource database will continue to expand its available set of resources and the user database is growing with increasing number of learners, thus forming a rich and exciting biomedical data science community.

ERuDIte: The Educational Resource Discovery Index for Data Science Learning

Jose-Luis Ambite, University of Southern California Information Sciences Institute; *Kristina Lerman, University of Southern California Information Sciences Institute; Lily Fierro, University of Southern California Information Sciences Institute; Jonathan Gordon, University of Southern California Information Sciences Institute; Florian Geigl, University of Southern California Information Sciences Institute; Knowledge Technologies Institute, Graz University of Technology; Gully A.P.C. Burns, University of Southern California Information Sciences Institute; John Darrell Van Horn, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Jeana Kamdar, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Sumiko Abe, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Avnish Bhattrai, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Xiaoxiao Lei, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Crystal Stewart, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute*

Data Science is a rapidly evolving field that draws techniques from many disciplines. There is a large number of resources available online for learning Data Science. However, these materials are highly heterogeneous, ranging from Massive Open Online Courses (MOOCs), to videos of tutorials and research talks presented at conferences, to books, blog posts, and standalone webpages. Consequently, any general search for “data science” will yield results ranging in difficulty, format, and topic, making the field intimidating to enter and difficult to navigate. In order to facilitate learning in Data Science, we are developing ERuDIte, the educational resource discovery index that powers the BD2K Training Coordinating Center Web Portal, which seeks to address the multiple issues involved in gathering and organizing heterogeneous learning resources. The development of ERuDIte has focused on three key areas: collection, integration, and organization of training resources. In the collection stage, we have manually identified sources of high-quality content, and we have built an automated web-scraping system to extract rich data from these sources. In the integration stage, we have designed a unified schema to integrate heterogeneous resource data into a single model that serves as a source for faceted search and may be expressed as linked data to facilitate information sharing. In the organization stage, we use methods from machine learning, information retrieval, and natural language processing to tag resources with concepts from a hierarchical, multi-dimensional taxonomy designed for the Data Science field. In summary, ERuDIte not only serves as a resource collector and aggregator, but also as a system powered by Data Science to intelligently organize resources and eventually provide a dynamic and personalized curriculum for biomedical researchers interested in learning about Data Science.

Training Component Activities at the Center for Causal Discovery

Takis Benos, Center for Causal Discovery

The goals of the training component of the Center for Causal Discovery (CCD) include the training of data scientists and biomedical investigators in the use of causal discovery tools and the dissemination of those tools. This year, the CCD held its second annual five-day summer short course in causal discovery from biomedical data. A total of 71 participants attended the course, including 33 graduate students, 5 postdoctoral fellows, 15 faculty, 10 undergraduate students, with the remaining 8 from industry, the healthcare sector, and government. A total of 21 different institutions were represented. The course included an introduction to graphical models, estimation and inference in causal models, model equivalence and search, and applications to a variety of biomedical big data problems including fMRI, chronic lung disease pathways, and cancer genomics. Following the short course, we held our first CCD Causal Discovery Datathon. A total of 34 participants attended including many who also attended the short course. Participants learned to use the CCD TETRAD software and to apply it to their own data. This past summer, we also inaugurated our undergraduate Big Data research experience in collaboration with BD2K R25 investigators at University of Puerto Rico (UPR). A total of 6 students from UPR spent ten weeks in Pittsburgh, working directly with CCD faculty and others on Big Data research projects. UPR undergraduates also attended part or all of the Summer Short Course as part of their training. Additional CCD Training Component activities included the Causal Discovery Distinguished Lecture Series, online curriculum development, and outreach activities at scientific meetings.

Preparing Underrepresented and First-Generation Students for Careers in Biomedical Big Data Science

Judith Canner, California State University, Monterey Bay; *Carla Fresquez, California State University, Monterey Bay*

In 2015, through the support of the NIH BD2K dR25 Enhancing Diversity in Biomedical Data Science Grant, California State University, Monterey Bay (CSUMB), a federally classified Hispanic Serving Institution, started to develop several programs to support undergraduate training in biomedical data science. The programs include 1) a summer research program for undergraduates at the Center for Big Data in Translational Genomics at UC Santa Cruz; 2) new majors, concentrations, and minors relevant to biomedical data science; and 3) professional and research training for undergraduate students and faculty at CSUMB. We will present the successes of the first year of the program and discuss plans for the upcoming years. Specifically, we will talk about the recruitment, training, and establishment of research experiences for 8 undergraduate students their ongoing professional development, and dissemination of their research. In addition, we will discuss how faculty at CSUMB in fields closely related to biomedical data science enhanced their training in order to conduct research with undergraduates and to teach/develop new courses relevant to biomedical data science. Finally, we will present new and planned curriculum development at CSUMB including an interdisciplinary statistics major and a concentration in bioinformatics. Though our current emphasis is translational genomics, due to our partnership with UCSC, we hope to extend our program to encompass applications of data science to other fields of biomedical research. In keeping with the CSUMB Vision of serving the diverse people of California, especially the working class and historically undereducated and low-income populations, the Biomedical Data Science Program at CSUMB will continue to evolve and grow over the next four years with focus on recruitment of native and transfer students into the program, dissemination of curriculum to other undergraduate minority serving institutions, and continued research experiences and training for students and faculty.

Biomedical Big Data Training for Novices: Initial Experience With a Short-Term Summer School

Brian Chapman, University of Utah; *Wendy W. Chapman, University of Utah; Karen Eilbeck, University of Utah; Matthew H. Samore, University of Utah; Donna Ziegenfuss, University of Utah*

We developed a short-term training program in biomedical data science to provide foundational training in biomedical data science, improve collaboration between clinicians and data scientists, and provide foundational knowledge for further biomedical data science education. We created an online learning environment based on Docker and Jupyter notebooks. Students access their learning environment via web browsers (no software installation). The Jupyter notebooks provide interactive programming environments and a bash emulator. We offered two nine-day courses (biomedical data science bootcamp and natural language processing), and a three-day workshop on time-series analysis. Using Python, the students learned programming, data visualization, statistical analysis, relational databases, machine learning, and domain standards within the context of data-centric problems from healthcare and biological research. Students also learned Git for version control. Guest lecturers described data centric research they were leading. Participants came from a wide range of backgrounds, including clinicians, librarians, biologists, informaticists, and computer scientists. To guide course revision we collected feedback from students daily as well as at the beginning and end of each course.

Students reported the value of the 'hands-on' learning environment integrated with expert guest lectures on related topics and, the ability to learn with and from each other. Interdisciplinary group-based exercises with just-in-time instructors and TA support helped students develop an understanding and comfort level with tools and computing environments. However, one of the biggest challenges identified in the short-term pilot course was the broad range of experience and knowledge that the participants brought to the course. Based on the formative feedback collected in the pilot summer course, future course iterations of the course will include more case-based exercises and supplemental technical documentation. Plans are also underway to design pre-course supplemental modules to help even out the computing and programming experience coming into the course.

Decaying Relevance of Clinical Data When Predicting Future Decisions

Jonathan Chen, Stanford University; *Muthuraman Alagappan, Internal Medicine Residency Program, Beth Israel Deaconess Medical Center; Mary K. Goldstein, Geriatrics Research Education and Clinical Center, Veterans Affairs Palo Alto Health Care System; Primary Care and Outcomes Research (PCOR), Stanford University; Steven M. Asch, Department of Medicine, Stanford University, Center for Innovation to Implementation, Veterans Affairs Palo Alto Health Care System; Russ B. Altman, Department of Medicine, Stanford University, Departments of Bioengineering and Genetics, Stanford Universit*

Objective

Determine how varying longitudinal historical training data can impact prediction of future clinical decisions. Estimate the "decay rate" of clinical data source relevance.

Materials and Methods

We trained a clinical order recommender system, analogous to Netflix or Amazon's "Customers who bought A also bought B…" product recommenders, based on a tertiary academic hospital's structured electronic health record data. We used this system to predict future (2013) admission orders based on different subsets of historical training data (2009 through 2012), relative to existing human-authored order sets.

Results

Predicting future (2013) inpatient orders is more accurate with models trained on just one month of recent (2012) data than with 12 months of older (2009) data (ROC AUC 0.91 vs. 0.88, precision 27% vs. 22%, recall 52% vs. 43%, all $P<10\text{-}10$). Algorithmically learned models from even the older (2009) data was still more effective than existing human-authored order sets (ROC AUC 0.81, precision 16% recall 35%). Training with more longitudinal data (2009-2012) was no better than using only the most recent (2012) data, unless applying a decaying weighting scheme with a "half-life" of data relevance about 4 months.

Discussion

Clinical practice patterns (automatically) learned from electronic health record data can vary substantially across years. Gold standards for clinical decision support are elusive moving targets, reinforcing the need for automated methods that can adapt to evolving information.

Conclusions and Relevance

Prioritizing small amounts of recent data is more effective than using larger amounts of older data towards future clinical predictions.

Training and Implementing Genomic Big Data Courses at Primarily Undergraduate Serving Institutions

Jeffrey Chuang, The Jackson Laboratory for Genomic Medicine; *Joel H. Graber, The Jackson Laboratory; Charles G. Wray, The Jackson Laboratory; Reinhard Laubenbacher, University of Connecticut Health Center*

Jackson Laboratory's Big Genomic Data Skills Training for Professors (Big Data) program was launched in May 2016. The Big Data program seeks to train college professors at undergraduate focused institutions to understand and subsequently teach big genomic data skills to their students. The inaugural training course was over-subscribed by 40%. College professors came from three Historically Black Colleges or Universities (HBCUs), four minority serving institutions (by U.S. Department of Education definition), and 15 INBRE institutions (NIH funded institutions within the Idea Network of Biomedical Research Excellence program). The inaugural training course covered a range of curricular topics including: RNAseq, single-cell RNAseq, using NGS for chromatin profiling, Whole Genome Seq/Mutation and Variant Analysis, Reproducibility and Robustness, and integrated modeling of heterogeneous data types. The JAX Big Data program is not a one-time course; rather it is a year-round effort that will support participants as they launch Big Genomic Data training with undergraduates at their home institutions. Working with participating professors it became clear that data analysis platform choice will be paramount to successful integration of genomic big data analysis within undergraduate courses. We have begun development of undergraduate-scoped genomic data training modules for professors' use. Currently three big data modules are being designed: High Throughput sequence QC, Expression profiling by RNAseq, and Variant discovery through genome sequencing in Human Disease.

Big Data Research and Education Program in a Primarily Undergraduate Institution (PUI)

Math Cuajungco, California State University, Fullerton*; Sam Behseta, California State University, Fullerton; Arthur W. Toga, University of Southern California; Sinjini Mitra, California State University, Fullerton; Harmanpreet Chauhan, California State University, Fullerton; Archana Jaiswal McEligot, California State University, Fullerton*

The expansion of Big Data science (BDS) has presented a compelling need to train a diverse biomedical workforce that has the ability to generate Big Data, as well as utilize and apply them in various fields of study. California State University, Fullerton (CSUF), a PUI classified as Hispanic-serving institution has partnered with the University of Southern California (USC), a Big Data for Discovery Science NIH BD2K Center of Excellence. Our NIH-funded Big Data Discovery and Diversity through Research Education Advancement and Partnerships (BD3-REAP) program aims to: (i) train and engage three cohorts of six predominantly under-represented students; (ii) train CSUF faculty on BDS research; and (iii) develop and integrate BDS curricula within and between departments at CSUF. For the first cohort, the program received a total of 22 applications (n=8 males, n=14 females), in which 59% of the applicants are first generation college students, and 82% are eligible for financial aid. The ethnic background of the applicants consists 45% Hispanic/Latino background, 5% African-American, 5% White, 31% Asian/Filipino-American, 9% Middle-Eastern, and 5% Other/Mixed race. The cohort will undertake independent research projects for two years, including a supplementary training in neuro-imaging and Big Data analytics at USC during the summer months. Meanwhile, curriculum development has been approved for implementation in spring 2017. Also, CSUF faculty along with USC collaborators have developed a thirty-minute instructional video complementing lecture materials for the integration of BDS in specific courses offered by the Colleges of Health and Human Development, and Natural Sciences and Mathematics. Pre- and post-treatment survey evaluation will reveal curricular effectiveness. Thus, full engagement of diverse students and faculty in BDS programs is highly feasible and a meritorious endeavor. The project aims are being accomplished according to the proposed timeline and additional progress is expected to be achieved in the coming months.

Biomedical Research Data Management Open Online Education: Challenges and Lessons Learned

Julie Goldman, University of Massachusetts Medical School; Elaine R. Martin, Harvard Medical School

The Best Practices for Biomedical Big Data project is a two year collaboration between Harvard Medical School and University of Massachusetts Medical School, funded by the NIH Big Data to Knowledge (BD2K) Initiative for Resource Development. The Best Practices for Biomedical Research Data Management Massive Open Online Course (MOOC) provides training to librarians, biomedical researchers, undergraduate and graduate biomedical students, and other interested individuals on recommended practices facilitating the discoverability, access, integrity, reuse value, privacy, security, and long term preservation of biomedical research data. This poster highlights lessons learned from the first year of this project.

Built upon the New England Collaborative Data Management Curriculum, the development team sought to use existing curricular materials to create a fully online course. The course is designed with an open course platform, WordPress Learning Management System (WPLMS), in order to facilitate broad access. Each of the MOOC's nine modules is dedicated to a specific component of data management best practices and includes video lectures, presentation slides, research teaching cases, readings, activities, and interactive quizzes.

The project team overcame multiple challenges related to creating an open online course: curriculum, audience and software.

Working towards overcoming these, the Best Practices for Biomedical Research Data Management MOOC development team has moved slowly and deliberately, created additional content, and added content experts to provide guidance. These lessons learned will assist course development beyond this project, adding to best practices for creating massive open online courseware. Lessons learned include: teaching method influences the curriculum and content should not be developed in isolation from the teaching method; content is dependent on audience and supplementary content can be used to bridge audience gaps; and implementing new or unfamiliar technologies is challenging so allow more time in the timeline for project team to work with open source platform.

Demystifying Biomedical Big Data: A Free Online Course

Yuiry Gusev, Georgetown University; *Bassem Haddad, Georgetown University Medical Center; Peter McGarvey, Georgetown University Medical Center*

Biomedical Big Data has become a sign of our times in this new genomics era, marked by a major paradigm shift in biomedical research and clinical practice. Advances in genomics have led to the generation of massive amounts of data. However, the usefulness of these data to the basic scientist or to the clinical researcher, to the physician or ultimately to the patient, is highly dependent on understanding its complexity and extracting relevant information about specific questions. The challenge is to facilitate the comprehension and analysis of big datasets and make them more "user friendly". Towards this end we developed a cross-disciplinary, Massive Open Online Course (MOOC), that aims to facilitate the understanding, analysis, and interpretation of biomedical big data for basic and clinical scientists, researchers, and librarians, with limited or no significant experience in bioinformatics. The 8-week course is funded by an NIH- BD2k R25 grant and, is planned to be released in January 2017 on the edX platform, where it will be freely available to anyone around the world. The course covers biomedical big data as it relates to five main areas: 1) Genomics; 2) Transcriptomics; 3) Proteomics; 4) Systems Biology; and 5) Big Data usage in translational research and the clinic. Content consists of short video lectures, interviews, and online hands-on-training on the use of various open source databases and analysis tools for different "omics" platforms. Online readings and resources will be provided to the students who can also participate in a discussion board. We plan to maintain this course as a "living resource," and update it regularly. We believe that this will allow us to provide an educational opportunity to a large audience worldwide, particularly to individuals with limited access to traditional educational resources in this cutting-edge field.

Data Science Education With MOOCs and Active Learning

Rafael Irizarry, Harvard University, Dana-Farber Cancer Institute*; Stephanie Hicks, Dana-Farber Cancer Institute*

Educational institutions across the world are responding to the unprecedented demand of training in statistics and data science by the creation of new courses, curriculums and degrees in applied statistics and data science. We have participated in two data science courses taught at Harvard and the creation of an online course of data analysis for the life sciences. In this presentation, I will discuss our approach to developing a MOOC based almost exclusively on real-world examples and how our lecturers revolved around dozens of exercises that required R programming to answer. We taught a total of seven different MOOCs in topics ranging from basic statistics to the analysis of RNA-Seq data. I will also discuss how the experience of teaching a MOOC changed the way I teach in the classroom. Specifically I will describe how we transformed the usual classroom by using active learning and collaborative techniques to teach concepts in statistics and data science. Examples include motivating real-world problems with data and code instead giving of traditional lectures, using Google Polls to get live feedback, and teaching the importance of reproducible research and collaborative practices with git/GitHub. In every assignment in the classroom, students performed a complete data analysis integrated programming skills with statistical analyses. As a final project, students analyzed a dataset of their choice and created a website and two minute video summarizing their results. This led to many students successfully obtaining jobs by discussing their homework and final projects in job interviews.

Community Training and Outreach Activities of the BD2K-LINCS DCIC

Sherry Jenkins, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Stephan Schürer, BD2K-LINCS DCIC, Center for Computational Science, University of Miami; Mario Medvedovic, BD2K-LINCS DCIC, Department of Environmental Health, University of Cincinnati College of Medicine; Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

The BD2K-LINCS DCIC Community Training and Outreach (CTO) component informs and educates key biomedical research communities about LINCS data and resources. The CTO efforts established several educational programs including a LINCS massive online open course (MOOC), a summer undergraduate research program, the initiation and support of diverse collaborative projects leveraging LINCS resources, and the systematic dissemination of LINCS data and tools via a variety of mechanisms. Recent activities of our CTO component include:

1. A LINCS-related MOOC: “Big Data Science with the BD2K-LINCS Data Coordination and Integration Center” https://www.coursera.org/learn/bd2k-lincs designed to train biomedical researchers with LINCS-related experimental methods, datasets, and computational tools.
2. Our second cohort of students for our BD2K-LINCS DCIC Summer Research Training Program in Biomedical Big Data Science. This is a research-intensive ten-week training program for undergraduate and graduate students.
3. Establishment of the webinar series “LINCS Data Science Research Webinars” which provides a forum for data scientists within and outside of the LINCS program to present their work on problems related to LINCS data analysis and integration. These webinars are posted on our YouTube channel at: https://www.youtube.com/playlist?list=PL0Bwuj8819U-G9Ob0jIGHp5AtwpCghLV5
4. Engagement in collaborative external data science research projects which focus on mining and integrating data generated by the LINCS program for new scientific discovery.
5. Hosting of symposia, seminars and workshops to bring together the DCIC and researchers who utilize LINCS resources.
6. The development and maintenance of the lincs-dcic.org and lincsproject.org websites as well as an active presence on various social media platforms including YouTube, Google+, and Twitter.

We established comprehensive education, outreach, and training programs aimed at scientific communities that can benefit from LINCS data and tools. We expect that the LINCS community will continue to grow into a resourceful network that brings together researchers across disciplines and organizations.

Community Research Education and Engagement for Data Science

Patricia Kovatch, Icahn School of Medicine at Mount Sinai; *Andrew Sharp; Luz Claudio*

Community Research Education and Engagement for Data Science (CREEDS) represents our commitment to our overall goal of fostering practical skills for a national, diverse and interdisciplinary community of early career researchers. Our three specific aims are: (1) to provide biomedical researchers with the practical skills and insight needed to harness the power and advance the promise of big data science to accelerate scientific discovery, (2) to develop an online social environment to facilitate the exchange of big data ideas, approaches and techniques between novices and experts, and (3) to enhance the diversity of the biomedical big data workforce through targeted recruitment and retention of disadvantaged and underrepresented student populations. We will personally engage 150 graduate students through an intensive, self-tailored, two-week summer school in NYC that will showcase interesting, current, collaborative case studies in activities at schools throughout NYC. Participants will employ active learning techniques to develop their skills of specific new methods and tools through both individual and group tasks on real-life large data sets. The training will raise the skills of students of varied backgrounds to a sufficient level for additional graduate research and will not require any prior computing experience. Students will also receive experience and materials to help them teach others when they return to their home institutions. Additionally, we will mentor another 30 NYC-based graduate students for team participation in four month long DREAM challenges. We will reach more individuals by placing the summer school on Coursera.

Get Real: A Synthetic Dataset Illustrating Clinical and Genetic Covariates

Ted Laderas, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; *Nicole Vasilevsky, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; Melissa Haendel, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; Shannon McWeeney, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; David A. Dorr, Department of Medical Informatics and Clinical Epidemiology and Department of Medicine, Oregon Health & Science University*

One challenge in training capable students in data science is the development of skills in the use of data, given the privacy and security concerns with patient-level data. Ideally, working with realistic, synthetic data can prepare the student for the numerous challenges in working with clinical and genetic data without the associated risks from actual patient data. We propose a 'solvable' and teachable analysis problem modeled on predicting high/low cardiovascular risk in a realistic patient cohort with both a genetic and a clinical component. To achieve this, we are developing a synthetic dataset based on a predetermined decision tree whose branches reflect various cardiovascular risk groups. These risk groups are based on combinations of covariates with known high cardiovascular risk (such as older patients with Type 2 Diabetes), but also "surprising" risk groups (for example, a younger patient with no smoking or diabetes, but who have a variant of a single nucleotide polymorphism (SNP) associated with CV risk). Frequencies of these risk groups in the overall cohort is derived from actual patient data, which determine sampling restrictions in the covariates associated with each risk group. By changing the associated probabilities of each risk group, the dataset can be "tuned" in levels of difficulty. We demonstrate the difficulty of the dataset on multiple machine learning methods, and we plan to release both the data model and the generation script in R and on GitHub. We believe that this dataset will give students valuable hands-on experience with working with both clinical and genetic data without the risks of working with identified or even de-identified data.

Getting Your Hands Dirty With Data

Ted Laderas, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; *Melissa Haendel, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; Nicole Vasilevsky, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; Bjorn Pederson, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; Jackie Wirz, Career and Professional Development Center, Oregon Health & Science University; William Hersh, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; David A. Dorr, Department of Medical Informatics and Clinical Epidemiology and Department of Medicine, Oregon Health & Science University; Shannon McWeeney, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University*

As part of the Big Data to Knowledge (BD2K) initiative, a research team at Oregon Health & Science University have developed a set of skills courses with the goal of training a diverse set of students in key issues related to data science. The courses include an introductory offering for undergraduate students, to more advanced graduate students, post-docs, research staff and faculty. To date, we have offered five in-person skills courses taught in a variety of ways, from a one week intensive, 2-4 evenings, and a half-day session. The topics are aimed to fill gaps in knowledge in topics such as problem description, data set discovery, preparation and creation of data sets, implementation of analytic techniques, ethics, biocuration, research data management and effective communication. Evaluations and feedback from the course were positive, with weaknesses in areas of engaging the students in discussion, providing a balance between instruction and practice, and creating content that is appropriate for the stated level of the class. When comparing the evaluations from beginner students to advanced students, we found the feedback was more positive from the advanced students. This may be because the advanced courses were more hands-on and application based, where students worked to solve specific computational problems, and the beginner material was primarily delivered as lectures and discussions and was less hands on. To address concerns, we are improving the hands-on components by including realistic synthetic datasets with clinical and genomics data that can be analyzed a variety of ways. Materials from the courses are shared on Figshare (bit.ly/DataAfterDarkDay1 and bit.ly/DataAfterDarkDay2) and more information is available here: ohsu.edu/bd2k.

Engaging and Training Undergraduates in Big Data Analysis Through Genome Annotation

Wilson Leung, Washington University in St. Louis; *Remi Marenco, The George Washington University; Yating Liu, Washington University in St. Louis; Jeremy Goecks, Oregon Health & Science University; Sarah C.R. Elgin, Washington University in St Louis*

As data science becomes increasingly important in biomedicine, it is critical to introduce students to "big data" early in their studies, to prepare them for jobs in industry and for graduate education. To meet the needs of introductory data science training, we are developing G-OnRamp, a suite of software and training materials that enables anyone new to big data analysis (e.g., undergraduates) to develop data science skills through eukaryotic genome annotation.

Genome annotation—identifying functional regions of a genome—requires use of diverse datasets and many algorithmic tools. Annotators must interpret potentially contradictory lines of evidence to produce gene models that are best supported by the available evidence. The Genomics Education Partnership (GEP; http://gep.wustl.edu) is a consortium of over 100 colleges and universities that provide classroom undergraduate research experiences in bioinformatics/genomics for students at all levels. The GEP is currently focused on the annotation of multiple Drosophila species. G-OnRamp will enable GEP faculty to diversify, using any eukaryote with a sequenced genome that fits their particular pedagogical and research interests.

G-OnRamp is a Galaxy workflow that creates a genome browser for a new genome assembly. Galaxy (http://galaxyproject.org/, https://usegalaxy.org) is an open-source, web-based scientific gateway for accessible, reproducible, and transparent analyses of large biomedical datasets that is used throughout the world. G-OnRamp extends Galaxy with (a) analysis workflows that create a graphical genome browser for annotation, including evidence from sequence homology, gene predictions, and RNA-seq, and (b) a stand-alone virtual machine to ensure wide availability. Future versions of G-OnRamp will include (i) interactive visual analytics; (ii) collaborative genome annotation; and (iii) a public server for broad usage. Concomitant with the development of the G-OnRamp software, we are also developing training materials that can be used by educators in an instructional setting and by individual researchers.

KnowEnG Tools for Barrier-Free Learning of Genomic Data Clustering

Mohith Manjunath, University of Illinois at Urbana-Champaign; *Yi Zhang, University of Illinois at Urbana-Champaign; Steve H. Yeo, University of Illinois at Urbana-Champaign; Omar Sobh, University of Illinois at Urbana-Champaign; Nathan Russell, University of Illinois at Urbana-Champaign; Christian Followell, University of Illinois at Urbana-Champaign; Colleen Bushell, University of Illinois at Urbana-Champaign; Umberto Ravaioli, University of Illinois at Urbana-Champaign; Jun S. Song, University of Illinois at Urbana-Champaign*

The advent of next-generation sequencing has enabled researchers to generate big data of various kinds at an unprecedented rapid pace. Therefore, there is acute need for resources that can enable the users to perform “first-hand” analysis, such as clustering, of high-dimensional data quickly. As part of the KnowEnG BD2K Center, we have developed a web-based resource called ClusterEnG (acronym for Clustering Engine for Genomics) for clustering big data with efficient parallel algorithms and software containerization. ClusterEnG offers a one-stop web service for clustering with the flexibility of choosing among many state-of-the-art clustering algorithms, which are not readily accessible to beginners. In order to facilitate the visualization of high-dimensional data, we implemented interactive versions of principal component analysis, a popular dimensional reduction technique, showing plots in 2D and 3D to allow intuitive exploration of structures in data. ClusterEnG also aims at educating the user about the similarities and differences between various clustering algorithms and provides clustering tutorials that demonstrate potential pitfalls of each algorithm.

Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K) in Puerto Rico

Patricia Ordóñez, University of Puerto Rico, Río Piedras; *José E. García-Arrarás, Department of Biology, University of Puerto Rico, Río Piedras; María E. Pérez, Department of Mathematics, University of Puerto Rico, Río Piedras; Humberto Ortiz-Zuazaga, Department of Computer Science, University of Puerto Rico, Río Piedras; Luis Pericchi, Department of Mathematics, University of Puerto Rico, Río Piedras*

Unprecedented advances in digital technology have produced a Big Data revolution that is transforming biomedical and clinical research. The IDI-BD2K Program at the University of Puerto Rico Río Piedras (UPR-RP) is focused on increasing the number of underrepresented researchers, both students and faculty, in Biomedical Big Data Science and increasing its application to biomedical research on the island. Four concrete aims were developed to achieve this goal:

1. Develop a biomedical big data curriculum,
2. Provide a cohort of undergraduate and graduate students with onsite research experiences in Interdisciplinary Biomedical Big Data projects,
3. Provide a cohort of at least 6 undergraduate students with summer research experiences at BD2K Centers every year, and
4. Offer opportunities for faculty and graduate students in Puerto Rico to train in areas of Big Data research, expand their expertise in Big Data, and thereby spur innovation in biomedical Big Data research on the island.

To achieve these aims, the UPR-RP partnered with researchers at BD2K Centers at Harvard University, the University of Pittsburgh, and the University of California Santa Cruz. This abstract describes the achievements of and the lessons learned in the first year of the implementation of the program.

Preparing Medical Librarians to Understand and Teach Research Data Management

Kevin Read, New York University School of Medicine; *Catherine Larson, New York University School of Medicine; Karen Yacobucci, New York University School of Medicine; So Young Oh, New York University School of Medicine; Suvam Paul, New York University School of Medicine; Colleen Gillespie, New York University School of Medicine; Alisa Surkis, New York University School of Medicine*

Two curricula were created to prepare medical librarians to teach research data management to researchers. The first is a web-based curriculum that uses interactive educational technologies to provide librarians with a better understanding of research data management, as well as the practice and culture of research. A second curriculum was designed as a toolkit to be used by librarians for in-person teaching of researchers and consists of slides, scripts, evaluation tools, and instructions. All educational materials were reviewed based on evidence-based instructional design and cognitive learning theories. Several modules in the toolkit allow for guided customization on topics where institution specific material is important (e.g. data storage). Evaluation of educational materials includes pre- and post-assessments of knowledge gain, satisfaction, comfort level with material, and intent to use. For the online modules, assessments are incorporated into the modules and for the teaching toolkit, a centralized online survey will be delivered before and after the in-person class. For both curricula, six month follow-up surveys will be conducted. Prior to broad dissemination, the online curricula was piloted to assess the quality of the modules and receive feedback from community. Future piloting will include visiting library sites and observing librarians teach researchers in-person. Following piloting, these curricula will be disseminated to the broader medical librarian community, for use at institutions across the United States to facilitate biomedical data management, sharing, and reuse.

MD2K Center of Excellence: Training and Development of a Transdisciplinary mHealth Workforce

Vivek Shetty, University of California, Los Angeles; Santosh Kumar, University of Memphis

The Mobile Sensor Data-to-Knowledge (MD2K) Center of Excellence has developed a wide variety of training resources to build and enable a sustainable community of transdisciplinary mHealth data scientists. First, MD2K organizes an annual, week-long immersive boot-camp in transdisciplinary mHealth methodologies and approaches. Co-funded by a R25 grant from OBSSR/NIDA, the mHealth Training Institutes (mHTI) held at UCLA has trained over 70 mHealth scholars from 50+ institutions to date. Selected from a very competitive pool of applicants, the mHTI scholars are drawn from across the academic spectrum. All scholars undergo pre– and post-mHTI assessments by a team of educational specialists at the UCLA Center for Research on Evaluation, Standards, and Student Testing. The qualitative and quantitative feedback from the scholars has been extremely positive and the feedback informs and refines future mHTI's.

Second, MD2K has created a virtual collaboratory called mHealthHUB, a dedicated website that serves as an organizing hub and online repository of mHealth tools, technologies and educational materials as well as a forum for the rapidly growing community of mHealth researchers across the globe to connect and collaborate. Recorded videos of all the mHTI lectures and regular MD2K webinars are curated on this site for broad dissemination.

Third, MD2K develops and releases training manuals and videos to accompany the release of MD2K's open-source software mHealth platforms (mCerebrum and Cerebral Cortex). The educational materials help end-users understand how to use the MD2K's software for collecting, curating, analyzing, and interpreting high-frequency data collected from various mobile sensors. Open-source code is made available at public GitHUB repositories for the community to adapt the software for their needs and contribute to refinements.

This talk will familiarize the BD2K community to the MD2K's training activates and resources, and stimulate discussion on how to use and contribute to MD2K's training resources.

The BD2K Training Coordinating Center: A Resource for the Data Science Community

***John Van Horn, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute**; Jeana Kamdar; Crystal Stewart; Sumiko Abe; Avnish Bhattrai; Xiaoxiao Lei; Krithika Ramaswamy; Shobhit Agarwal; Carmen Tan; Caroline O'Driscoll; Lily Fierro; Jose-Luis Ambite; Kristina Lerman; Jonathan Gordon; Gully Burns; Rochelle Tractenberg; Florian Geigl; Priya Jain; John Berardo; Michael Taylor*

The BD2K Training Coordinating Center (TCC) helps to promote and support training and educational activities across the collection of NIH funded Big Data to Knowledge (BD2K) programs.[1] In particular, the TCC brings an innovative approach to the exploration and discovery of biomedical Big Data educational and training content. The TCC and its Big Data U website (bigdatau.org) aim to equip biomedical researchers with the tools and techniques to navigate and interpret Big Data accurately, efficiently, and effectively by providing a centralized and personalized online educational platform. The Educational Resource Discovery Index (ERuDIte) and 'Knowledge Map' serve as the engine of the Big Data U web platform, hosting an array of diverse domains, ranging from molecular biology to probability statistics. Big Data U functions as a comprehensive training resource that allows users to easily identify and engage with educational resources aligned with their experience and personal learning needs. Being a multifaceted center, the TCC also runs other various events and projects including the data science rotation RoAD-Trip program, an annual Data Science Innovation Lab series, the online Data Science Seminar Series (with the BD2K CCC and the NIH), is creating original films on the science behind 'big data' science, as well as contributes to numerous community outreach activities. All in all, through our BD2K TCC activities, these diverse projects are helping to support the rich network of data scientists and researchers and their desire to prepare, learn, analyze, and interpret large-scale biomedical data.

1. Van Horn, J. D. Opinion: Big data biomedicine offers big higher education opportunities. Proceedings of the National Academy of Sciences 113, 6322-6324, doi:10.1073/pnas.1607582113 (2016).

Data Science Educational Resources for Anyone, Anywhere

Nicole Vasilevsky, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; *Bjorn Pederson, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; Jackie Wirz, Career and Professional Development Center, Oregon Health & Science University; Shannon McWeeney, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University; Melissa Haendel, Department of Medical Informatics and Clinical Epidemiology and Library, Oregon Health & Science University; William Hersh, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University*

A research team at Oregon Health & Science University is developing freely available, online Open Educational Resources (OERs) that cover various topics in data science. The OERs can be used as 'out of the box' courses for students, or materials for educators to use in courses, training programs, or seminars. To inform development of our materials, we performed a needs assessment among leaders in Oregon undergraduate institutions to determine the gaps in knowledge and areas of interest. Additionally, we prioritized topics to fill gaps in the BD2K educational and training offerings. Our goal is to create 32 modules on topics including data standards, ethics, team science, secondary reuse of data, information retrieval, biocuration, ontologies, metadata, and data visualization. We also provide a mapping to research competencies in other areas, such as for the CTSA consortium research competency requirements, for general data science from the Oceans of Data initiative, and the Medical Library Association professional competencies for health sciences librarians. To this end, we are able to link these materials to existing efforts, and provide training opportunities for learners and educators working in these areas. We would like to complete this mapping across all of the BD2K training offerings, to align with other groups, avoid redundancy and to ensure we are meeting the needs of these various groups. The OERs are intended to be flexible and customizable and we encourage others to use or repurpose these materials for training, workshops and professional development or for dissemination to instructors in various fields. The OERs and other materials are available here: dmice.ohsu.edu/bd2k.

University of Washington's R25 Short Course

Daniela Witten, University of Washington

The Summer Institute in Statistics for Big Data consists of a series of 2½ day workshops (modules) designed to introduce biologists, quantitative scientists, and statisticians to modern statistical techniques for the analysis of Biological Big Data. The format will involve formal lectures, computing labs, and hands-on case studies.

Pandem-Data: Using Big Data in High School

Chuck Wood, Wheeling Jesuit University

The biomedical world is immersed in a continuing flood of information called big data. This deluge of data comes from every aspect of medical research, patient treatment and life science investigations. Big data overwhelms traditional means of data collection, storage and analysis, causing new fields to emerge to manage and exploit it. Teams of bioscientists, mathematicians, statisticians and computer scientists are all needed to extract useful results from big data, but an educational pipeline has yet to be developed to supply the needed expertise. Pandem-Data will use the current excitement and interest in epidemics and pandemics to engage high school students in the use of big data. Pandem-Data is an extension of Wheeling Jesuit University's NIH/SEPA Pandem-Sim project that immerses students within various epidemiological roles to investigate disease outbreaks.

Pandem-Data will instruct high school students in the use of modeling software to investigate the spread of infectious diseases and to evaluate the efficacy of strategies to reduce the impact of an outbreak. Students will learn about the power of big data through manipulation of disease variables including vaccination, school closings, and transportation impacts on an outbreak.

Students will be guided through open-ended explorations by the use of inquiry-based learning modules that will systematically introduce big data concepts, epidemiology, disease modeling, and the use of the software to build models and interpret their results. A comprehensive online professional development package will prepare teachers to use confidently the Pandem-Data modules in their classrooms.

Pandem-Data will employ a developmental evaluation approach to meet the project's rapid response requirements and to show promise for pre-formative design of a potentially broad-impact, scalable innovation. We expect Pandem-Data will be used in the team-teaching of epidemiology and computer science courses to inspire high school students to consider careers in big data science.

Big Data Training for Translational Omics Research

***Min Zhang, Purdue University**; James Fleet, Purdue University; Wanqing Liu, Purdue University; Pete Pascuzzi, Purdue University*

The educational goals of our BD2K funded boot camp are to raise participants' awareness and knowledge of the value of big data in biomedical research, build basic competency of participants in the use of established tools and public databases, and provide them a vocabulary to effectively communicate with big data science experts. During the first two-week summer boot camp held at Purdue University this year, we introduced the basic computational skills including R and Unix, demonstrated how to use public databases and tools, and analyzed different types of omics data. In addition, the utility of big data in biomarker discovery was discussed. Hands-on activities were complemented with formal lectures by guest speakers on various topics important for big data science, such as "Next Generation Sequencing", "The power of Statistics and Bioinformatics for helping to Unlock Epigenetic Secrets", and "Principles of Functional Neuroimaging", among many others. The workshop was well received by 30 participants from all over the country and significant improvements in working with biomedical big data were observed from comparing the pre- and post-course survey. Detailed information about the workshop can be found at http://www.stat.purdue.edu/bigtap/.

BioCADDIE & Resource Indexing

Aztec: A Cloud-Based Computational Platform to Integrate Biomedical Resources

Brian Bleakley, HeartBD2K Center at UCLA; *Chelsea Ju, HeartBD2K Center at UCLA; Vincent Kyi, HeartBD2K Center at UCLA; Justin Wood, HeartBD2K Center at UCLA; Patrick Tan, HeartBD2K Center at UCLA; Giuseppe Mazzeo, HeartBD2K Center at UCLA; Howard Choi, HeartBD2K Center at UCLA; Wei Wang, HeartBD2K Center at UCLA; Peipei Ping, HeartBD2K Center at UCLA*

Introduction

Omics phenotyping has become increasingly recognized in our path to precision medicine. A major computational challenge of our investigator community is to identify the necessary data analytical tools to process multi-omics data. To aid navigation of the analytical tools, we have created a novel computational resource platform, Aztec, that empowers users to simultaneously search a diverse array of digital resources including databases, standalone software, web services, publications, and large libraries composed of many interrelated functions.

Method

Aztec is cloud-based and is hosted on an Amazon EC2 server, with flexibility to scale up to meet user demand. Aztec is implemented on a Node.js framework, using an enterprise established Model-View-Controller design pattern. It employs a user management system with authentication and encryption to enable secure resource registration and updates. Aztec offers an advanced yet user-friendly query system, providing a cascade of nested sub-searches as well as comprehensive filtering and sorting of results. To preserve data consistency and integrity across diverse and voluminous inputs, Aztec employs a relational database implemented in MySQL. Powered by Apache Solr, Aztec supports, in one unified platform, efficient keyword search, and semantic query guided by established ontologies.

Results and Summary

On Aztec.bio there are currently 9,000+ resources spanning domains from imaging, gene ontology, text-mining, data visualization and various omics analyses, including genomics, transcriptomics, proteomics, and metabolomics. Aztec supports the biomedical community by ensuring the relevant resources remain findable and accessible, capturing the accomplishments of researchers and tool developers in a sustainable manner.

The bioCADDIE Data Citation Implementation Pilot

Tim Clark, Massachusetts General Hospital and Harvard Medical School; *Helena Cousijn, Elsevier B.V.; Mercè Crosas, Harvard University; Martin Fenner, DataCite; Jeffrey S. Grethe, University of California, San Diego; Nick Juty, European Bioinformatics Institute; Amye Kenall, SpringerNature; John Kunze, California Digital Library; Joan Starr, California Digital Library; Maryann Martone, University of California, San Diego*

The bioCADDIE Data Citation Implementation Pilot (DCIP) is an effort to speed development of coordinated data archiving and citation practices and associated infrastructure in biomedical research by working directly with major publishers, repositories, identifier/metadata registries, and identifier resolution services.

Our goal is to enable common data citation practices across the biomedical ecosystem consistent with the Joint Declaration of Data Citation Principles (JDDCP). We are working to eliminate obstacles, clarify open questions, and converge implementation plans through a series of workshops and telecons. We expect to release draft data citation roadmaps for publishers and repositories, with online FAQs; and to implement alignment of the major US and European identifier resolution systems using a common namespace/prefix registry, by the end of 2016.

DCIP was organized by bioCADDIE as a community group activity within FORCE11.org, to help ensure maximum input and involvement from ecosystem participants across the biomedical research, informatics, repositories and publishing communities.

Our presentation will provide an outline of the key results contained in these roadmaps and specifications. These include:

- Recommendations and implementation plans for identifiers developed in conjunction with the California Digital Library, the European Bioinformatics Institute, ELIXIR, NIH, and DataCite that build on recommendations from bioCADDIE's identifier working group.
- Specific mechanisms for PREFIX:ACCESSION based identifier resolution and common prefix registry maintenance.
- Publication lifecycle recommendations for author instructions, editorial training, approved repository list maintenance, and reference presentation.
- Repository landing page metadata serialization in human and machine readable form with both backward and forward ecosystem compatibility.
- Citation metadata alignment with bioCADDIE's DATS, and with DataCite, Dublin Core, and Schema.org.

Reactome: A Curated Knowledge Base of Biomolecular Pathways

Antonio Fabregat, EMBL-EBI; *Konstantinos Sidiropoulos, EMBL-EBI; Guilherme Viteri, EMBL-EBI; Florent Yvon, EMBL-EBI; Corina Duenas, EMBL-EBI; Steven Jupe, EMBL-EBI; Peter D'Eustachio, New York University School of Medicine; Lincoln Stein, Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory, Department of Molecular Genetics, University of Toronto; Henning Hermjakob, EMBL-EBI and National Center for Protein Sciences*

Reactome (http://www.reactome.org) is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways, aiming to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge and to support basic research, genome analysis, modelling, systems biology and education. Pathways are built from connected "reactions" that encompass many types of biochemical events. Reactions are derived from literature and must cite a publication that experimentally validates them. Pathways are authored by expert biologists and peer reviewed before incorporation into the database. In its latest release (v58), Reactome includes 10,168 reactions covering 10,212 human gene products and supported by 24,968 literature references. Users can search for proteins or compounds and see details of the complexes, reactions and pathways they participate in. Pathway diagrams allow users to examine the molecular events that constitute the steps in pathways and to view details of the proteins, complexes and compounds involved. Different forms of pathways analysis can be performed with the Reactome analysis tools. Users can submit a list of identifiers for overrepresentation analysis or submit quantitative datasets, such as microarray data, for expression analysis. Results of these analyses are overlaid onto the Pathways Overview and Diagram Viewer for easy navigation and interpretation. Interaction data from multiple resources can be used to expand pathways. Interactors from IntAct are included by default in the search feature and can be taken into account in the analysis service. Finally, pathways or all Reactome content can be downloaded in many formats including TSV, CSV, PDF, SBML, BioPax and PSI-MITAB.

SATORI: A System for Ontology-Guided Visual Exploration of Biomedical Data Repositories

Nils Gehlenborg, Harvard Medical School; *Fritz Lekschas, Harvard School of Engineering and Applied Sciences*

The ever-increasing number of biomedical data sets provides tremendous opportunities for re-use across multiple studies. In some cases individual published data sets can be used to test a hypothesis instead of generating new data. Alternatively, data from previous studies can be employed as corroborating evidence for observations made in an experiment. Meta studies that include data from dozens or hundreds of published data sets are another frequent use case for the re-purposing of previously generated data.

Current data repositories, however, provide limited means of exploration apart from text-based search. Ontological metadata annotations provide context by semantically relating data sets. Visualizing this rich network of relationships can improve the explorability of large data repositories and help researchers find data sets of interest and make biomedical data findable - the first guiding principle for FAIR (findable, accessible, interoperable, and reusable) research data.

To address the needs for precise search and exploration, we propose a system that is based on free text and ontologically-annotated metadata and consists of two highly interlinked interfaces: a powerful text-based search and a visual analytics exploration tool. We have designed and implemented SATORI - an integrative search and exploration interface for the exploration of biomedical data repositories. Our tool enables researchers to seamlessly browse and semantically query ontologically annotated repositories via two visualizations that are highly interconnected with a powerful search interface.

SATORI is integrated into the Refinery Platform which is open-sourced and freely available at http://refinery-platform.org. Additional information about SATORI is available at http://satori.refinery-platform.org.

A Framework for Metadata Management and Automated Discovery for Heterogeneous Data Integration

Ramkiran Gouripeddi, Department of Biomedical Informatics and Center for Clinical and Translational Science, University of Utah; *Peter Mo, Center for Clinical and Translational Science, University of Utah; Randy Madsen, Center for Clinical and Translational Science, University of Utah; Phillip Warner, Center for Clinical and Translational Science, University of Utah; Ryan Butcher, Center for Clinical and Translational Science, University of Utah; Jingran Wen, Department of Biomedical Informatics, University of Utah; Jianyin Shao, Department of Biomedical Informatics, University of Utah; Nicole Burnett, Department of Biomedical Informatics, University of Utah; Naresh Sundar Rajan, Department of Biomedical Informatics and Center for Clinical and Translational Science, University of Utah; Bernie LaSalle, Department of Biomedical Informatics and Center for Clinical and Translational Science, University of Utah; Julio C. Facelli, Department of Biomedical Informatics and Center for Clinical and Translational Science, University of Utah*

Current approaches to metadata discovery are dependent on time consuming manual curations. To realize the full potential of Big Data technologies in biomedicine, enhance research reproducibility and increase efficiency in translational sciences it is critical to develop automatic and/or semiautomatic metadata discovery methods and the corresponding infrastructure to deploy and maintain these tools and their outputs.

Towards such a discovery infrastructure: We conceptually designed a process workflow for Metadata Discovery and Mapping Service, for automated metadata discovery. Based on steps taken by human experts in discovering and mapping metadata from various biomedical data, we designed a framework for automation. It consists of a 3-step process: (1) identification of data file source and format, (2) followed by detailed metadata characterization based on (1), and (3) characterization of the file in relation to other files to support harmonization of content as needed for data integration. The framework discovers and leverages administrative, structural, descriptive and semantic metadata, and consists of metadata and semantic mappers, along with uncertainty characterization and provision of expert review. As next steps we will develop and evaluate this framework using workflow platforms (e.g. Swift, Pegasus).

In order to store discovered metadata about digital objects, we enhanced OpenFurther's Metadata Repository (MDR). We configured the bioCADDIE metadata specifications (Data Tag Suite (DATS) model) as assets in the MDR for harmonizing metadata of individual datasets (e.g. different protein files) for data integration. This method of metadata management provides a flexible data resource metadata storage system that supports versioning metadata (e.g. DATS 1.0 to 2.1) and data files mapped to different versions, enhance descriptors of resources (DATS) with descriptions of content within resources, and translations to other metadata specifications (e.g. schema.org). Also, this MDR stored metadata is available for various data services including data integration.

A Machine Learning Approach for Data Source and Type Identification to Support Metadata Discovery

Ramkiran Gouripeddi, Department of Biomedical Informatics and Clinical and Translational Science, University of Utah; *Jingran Wen, Department of Biomedical Informatics, University of Utah; Julio C. Facelli, Department of Biomedical Informatics and Clinical and Translational Science, University of Utah*

Current approaches to metadata discovery are dependent on time consuming manual curations and it is critical to develop automatic and/or semiautomatic metadata discovery methods to realize the full potential of Big Data technologies in biomedicine, enhance research reproducibility and increase efficiency in translational sciences. We are developing a two-step metadata discovery workflow: (1) Identification of data source and type using their intrinsic document structure, and (2) Discovery of detailed metadata within the data by associating specific metadata discovery tools based on the data's source and type ascertained from (1). Here we discuss our initial results for (1) using machine learning.

In this work we included biomedical data from different sources and file formats: Human genetic variants - ClinVar (XML, tab-delimited), Protein structure (PDB, mmCIF, XML), Biomedical literature, and General English corpus. We tokenized the data files using Natural Language Toolkit and considered various document structural features: normalized count of numerical tokens, negative numerical tokens, number of words, number of capitalized words, number of words with all upper letters, and median length of tokens. We developed a decision tree model with these structural features to classify these data and types, and evaluated the performance of the model using 10-fold cross-validation and test sets for the following metrics: precision, recall, and F1 score using scikit-learn. Our model was able to distinguish protein structure, genetic variant, scientific paper and general English files with an average F1 score of 0.997, 0.997, 0.886 and 0.919 when evaluated using cross-validation, and 1, 0.999, 0.980, and 0.935 when using independent test sets. Our approach shows it is possible to automatically identify data sources and types using only document structural features and therefore reasonable to programmatically associate metadata extraction tools specific for each data source and type as next steps.

A Scalable Dataset Indexing Infrastructure for the bioCADDIE Data Discovery System

Jeffrey Grethe, University of California, San Diego; *Burak Ozyurt, University of California, San Diego; Hua Xu, University of Texas Health Science Center at Houston; Xiaoling Chen, University of Texas Health Science Center at Houston; Ruiling Liu, University of Texas Health Science Center at Houston; Anupama Gururaj, University of Texas Health Science Center at Houston; Hyeon-eui Kim, University of California, San Diego; Yueling Li, University of California, San Diego; Nansu Zhong, University of California, San Diego; Claudiu Farcas, University of California, San Diego; Alejandra Gonzalez-Beltran, University of Oxford; Philippe Rocca-Serra, University of Oxford; Ergin Soysal, University of Texas Health Science Center at Houston; Ian Fore, National Institutes of Health; Ronald Margolis, National Institutes of Health; George Alter, University of Michigan, Ann Arbor; Susanna-Assunta Sansone, University of Oxford; Lucila Ohno-Machado, University of California, San Diego*

Introduction

The scalable dataset indexing infrastructure of the DataMed prototype maps the disparate metadata from the diverse data sources into a unified specification provided by various working groups organized by bioCADDIE and related communities. This pipeline involves an automated component that provides controlled translation and curation of metadata using special tools such as a transformation language and JSON-Path.

Overview of the bioCADDIE Dataset Indexing Infrastructure

Since many indexing projects already exist and many repositories are already maintaining detailed metadata about the data sets they host, our goal was to leverage all this work and build a cross-repository/cross-indices index. The overall infrastructure consists of the following components:

- A data and metadata extraction system that is able to connect to various repositories and data aggregators. All metadata information is converted to JSON documents for each dataset being described and stored in MongoDB.
- A messaging infrastructure, utilizing Apache ActiveMQ2, distributes dataset description documents from MongoDB, and depending on their status value, dispatches them to persistent point-to-point queues.
- A collection of multiple concurrent consumers retrieve the documents from MongoDB3, process it, update the job status and save it back to the MongoDB. Consumers can be written using the STOMP4 protocol. Documents are transformed, to align with the bioCADDIE metadata model, and processed by a collection of modules that enhance the metadata records.
- Fully processed documents are then exported to an ElasticSearch5 endpoint that serves the dataset indices via standard ElasticSearch RESTful services.

Current Status

With the latest release of the DATS metadata standard (v2.1) work is focused on incorporating new repositories to expand the content as well as on validating, testing and improving the metadata descriptions.

Acknowledgments

bioCADDIE has many more team members and collaborators than would fit the author list here. They are listed at http://biocaddie.org.

Deep Learning-Based Multi-Modal Indexing of Heterogeneous Clinical Data for Patient Cohort Retrieval

Sanda Harabagiu, University of Texas at Dallas; *Travis Goodwin, University of Texas at Dallas*

Heterogeneity is one of the attributes of clinical Big Data, as the clinical picture of patients is documented by narratives, images, signals, etc. The ability to search through such data is made possible by multi-modal indexes, capturing the knowledge processed across all forms of clinical documents. Taking advantage of state-of-the-art deep learning methods, we have been able to generate a multi-modal index operating on a vast collection of electroencephalography (EEG) signal recordings and EEG reports and employ it is a patient cohort retrieval system. When EEG reports were indexed, the sections of the EEG reports were identified and medical language processing was performed. When the EEG signal recordings were processed, they were represented by EEG signal fingerprints as low-dimensional vectors produced by deep learning methods on the Big Data of EEG signals. Moreover, we organize the EEG fingerprints into a similarity-based hierarchy, which was included in the multi-modal index.

We used the multi-modal index in a patient cohort retrieval system, called MERCuRY (Multi-modal EncephalogRam patient Cohort discoveRY), which relies on medical language processing to identify the inclusion and exclusion criteria from the queries generated by neurologists. Using state-of-the-art relevance models adapted for a multi-modal index, we obtained very promising results, indicating that the multi-modal index is bridging the gaps between the electrode potentials recorded in the EEG signal and the clinical findings documented in the EEG reports.

Omics Discovery Index – Discovering and Linking Public 'Omics' Datasets

Henning Hermjakob, EMBL-EBI; *Yasset Perez-Riverol, EMBL-EBI; Mingze Bai, EMBL-EBI and National Center for Protein Sciences; Gaurhari Dass, EMBL-EBI; Rodrigo Lopez, EMBL-EBI; Peipei Ping, Departments of Physiology, Medicine, and Bioinformatics, University of California, Los Angeles*

Biomedical data, in particular omics datasets are being generated at an unprecedented rate. As a result, the number of deposited datasets in public repositories originating from various omics approaches has increased dramatically in recent years. However, this also means that discovery of all relevant datasets for a given scientific question is non-trivial. Here, we introduce the Omics Discovery Index (OmicsDI), an integrated and open source platform facilitating the access and dissemination of omics datasets. OmicsDI provides a unique infrastructure to integrate datasets coming from multiple omics studies, including at present proteomics, genomics transcriptomics, and metabolomics, as a globally distributed resource.

As of October 2016, OmicsDI provides a lightweight discovery tool including more than 80,000 omics datasets from ten different repositories, four different omics types, and three continents. While advanced metadata-based browsing and indexing supports dataset findability, the lightweight approach avoids the development of redundant concepts and infrastructure. The original datasets are not replicated, but referenced. In the interest of sustainability, the responsibility for provision of a well-formatted metadata records lies with the original data providers, similarly to the concept of publisher data provision to PubMed or EuroPMC.

OmicsDI supports full text search as well as ontology-based search extensions. In addition, we use the concept of biological dataset similarity, based on the number of shared biological entities, for example protein identifications, among datasets. This allows us to suggest potential relationships among datasets even if they don't share sufficient metadata annotation.

OmicsDI is accessible at http://omicsdi.org.

The LINCS Data Portal and FAIR LINCS Dataset Landing Pages

Amar Koleti, Center for Computational Science, University of Miami; *Raymond Terryn, Center for Computational Science, University of Miami; Vasileios Stathias, Center for Computational Science, University of Miami; Michele Forlin, Center for Computational Science, University of Miami; Dušica Vidović, Center for Computational Science, University of Miami; Caty Chung, Center for Computational Science, University of Miami; Wen Niu, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, and BD2K LINCS Data Coordination and Integration Center; Caroline Monteiro, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, and BD2K LINCS Data Coordination and Integration Center; Christopher Mader, Center for Computational Science, University of Miami, and BD2K LINCS Data Coordination and Integration Center; Avi Ma'ayan, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, and BD2K LINCS Data Coordination and Integration Center; Mario Medvedovic, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, and BD2K LINCS Data Coordination and Integration Center; Stephan Schürer, Center for Computational Science, University of Miami; Department of Molecular and Cellular Pharmacology, University of Miami, and BD2K LINCS Data Coordination and Integration Center*

The LINCS Data Portal (LDP) presents a unified interface to access LINCS datasets and metadata with mappings to several external resources. LDP provides various options to explore, query, and download LINCS dataset packages and reagents that have been described using the LINCS metadata standards.

We recently introduced LINCS Dataset Landing Pages to provide integrated access to important content for each LINCS dataset. The landing pages provide deep metadata for each LINCS dataset including description of the assays, authors, data analysis pipelines, and standardized reagents such as small molecules cell lines, antibodies, etc, with rich annotations. The landing pages are a key component to make LINCS data persistent and reusable, by integrating LINCS datasets, data processing pipelines, analytes, perturbations, model systems and related concepts as uniquely identifiable digital research objects.

LDP supports ontology-driven concept search, free text search, facet filtering, logical intersection of filters (AND, OR), and list, table, and matrix views. LDP enables download of LINCS dataset packages, which consist of released datasets and associated metadata. LDP also provides several specialized apps including small molecule compounds and cell lines. A landing page facilitates interactive exploration of all LINCS datasets via several classifications.

LDP is built on a robust API and is integrated with the MetaData Registry and interfaces with other components of the Integrated Knowledge Environment (IKE) developed in our Center. All LINCS datasets are also indexed in bioCADDIE DataMed.

Augmenting the Capabilities for Semantic Search of the Medical Literature

Ani Nenkova, University of Pennsylvania; *Ari Brooks, University of Pennsylvania; Zachary Ives, University of Pennsylvania; Byron C. Wallace, Northeastern University*

Searching the medical literature, both for the purpose of producing systematic reviews and for patient-tailored clinical decisions, would benefit from interfaces that support browsing and discovery in addition to search for specific queries. For example a user may enter a medical condition and receive a result displaying what treatments have been studied for the condition, in which special groups and which outcomes have been used to assess effectiveness. This faceted interface would allow users to reformulate and sharpen their query more transparently than using trial and error approach based on impressions from a list of unstructured results.

We are developing automated methods for extracting the structured representation of the medical literature necessary to power such an interface. The target for extraction includes main medical condition, co-morbidities, treatment, and outcome measures. We aim to leverage noisy annotations obtained from crowdsourcing, patients and other lay volunteers, medical students and doctors.

To study the generalizability of our annotation scheme and the robustness of the automated extraction methods, we develop a corpus of 5,000 annotated abstracts of medical articles from five different domains. All abstracts are of randomized control trials. A fifth are drawn randomly from PubMed. Others are drawn from articles related to cancer, cardiovascular disease, autism and dementia.

In work done prior to the grant, the project PI has developed and deployed a system for detecting clinically relevant sentences in full text medical articles. The system, RobotReviewer, also detects risk of bias according to criteria used in the preparation of systematic reviews. The system can be accessed at http://www.robotreviewer.net/.

bioCADDIE: Progress and Next Steps

Lucila Ohno-Machado, University of California, San Diego; *Jeff Grethe, University of California, San Diego; Ian Fore, National Institutes of Health; Susanna-Assunta Sansone, University of Oxford; George Alter, University of Michigan; Hua Xu, University of Texas*

The bioCADDIE Data Discovery Index Consortium project started effectively in March 2015. It has three goals: help users find data, build a data discovery index, and interoperate with other entities in the NIH Commons. During the past 19 months, the recommendations of working groups, metadata specifications, software development, and pilot projects helped develop the first prototype of the DataMed search engine, available at datamed.org. We will report on our various engagement activities, summarize the collaborations with BD2K initiatives and other groups, and explain our vision and future directions for bioCADDIE.

Our prototype development involves work in several specific areas, including metadata management, metadata mapping, user-interface design and backend database development, which form the core development team. Additionally, working groups within bioCADDIE address specific needs within the consortium. Currently active working groups focus on Evaluation and on Outreach, among others. Together these groups worked together to release the DataMed v1.0 prototype, which went live on June 30, 2016 and contains 23 indexed repositories. Version 1.5 of DataMed will be available in late November, 2016 and will include an additional 18 repositories.

bioCADDIE also has many external collaborations, including with centers with an important digital object indexing interest, like Heart2BD2K, CEDAR, LINCS, Force11, and ELIXIR, among others. These collaborations serve an important role by allowing dedicated efforts towards sharing of ideas and goals, as well as cross-project integration of tools and resources.

Future developments for bioCADDIE include ingestion of many more repositories, and version 2.0 of DataMed release in February 2017. The currently ongoing Dataset Retrieval Challenge, which has 28 groups participating, will grant at least one group an opportunity to have their methods integrated into the DataMed tool. We will fund additional pilot project requests for applications to continue to connect with the community at large.

The DataMed DATS Model, Annotated With Schema.org

Susanna-Assunta Sansone, University of Oxford; *Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Mary Vardigan, Jeffrey Grethe, Hua Xu, and Members of the DataMed Development Team and of the bioCADDIE Working Groups*

The NIH BD2K bioCADDIE's DataMed stores metadata generic enough to describe any dataset using a model we have called the DAta Tag Suite (DATS). Akin to the Journal Article Tag Suite (JATS) used in PubMed, DATS enables submission of data for “ingestion” by DataMed.

DATS is a community-driven model designed to cover both (i) experimental datasets, which do not change after deposit in a repository, and (ii) datasets in reference knowledge bases describing dynamic concepts, whose definition morphs over time. The DATS model has a core and extended set of elements, to progressively accommodate more specialized data types. Like the JATS, the core elements are generic and applicable to any type of datasets. The extended DATS includes an initial set of elements, some of which are specific for life and biomedical science domains and can be further extended. The DATS entities are available as machine-readable JSON schemata, with examples and schema.org annotated JSON-LD serialization (DOI: 10.5281/zenodo.62024 and https://github.com/biocaddie)

The DATS model was developed given the following considerations:

- A variety of data discovery initiatives exists or are being developed, different scope, use cases and approaches. Their metadata schemas are valuable and were reviewed to determine essential items. The metadata schemas and models used in the mapping have been described in the BioSharing Collection (https://biosharing.org/collection/bioCADDIE)
- Identification of the initial set of metadata elements was based on: (i) analyses of use cases (a top-down approach); and (ii) mapping of existing metadata schemas (a bottom-up approach). From the use cases, a set of ‘competency questions’ were derived - these defined the questions that we want DataMed to answer- abstracted, key concepts binned in entities, attributes and values categories, to be easily matched with the results of the bottom-up approach.

Reactome: New Services and Widgets to Ease Third-Party Integration

Konstantinos Sidiropoulos, EMBL-EBI; *Antonio Fabregat, EMBL-EBI; Guilherme Viteri, EMBL-EBI; Florian Korninger, EMBL-EBI; Peter D'Eustachio, New York University School of Medicine; Lincoln Stein, Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory, Department of Molecular Genetics, University of Toronto; Henning Hermjakob, EMBL-EBI, National Center for Protein Sciences, Beijing*

Reactome (http://www.reactome.org) is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways. It aims to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modelling, systems biology and education. Thus, the mainstays of its software development are usability and responsiveness from the user's point of view, likewise modularity and reusability from the developer's side. Reactome offers web services and widgets (http://goo.gl/koRvhp) to facilitate integration in third-party software. One service provides database access while the other performs overrepresentation and expression analysis as well as species comparison. Widgets for the Pathways Overview and Pathway Diagrams are provided for JavaScript and GWT. Both widgets overlay the results of the Analysis Service. Protein-protein or protein-chemical interactions can be used to extend pathways beyond Reactome's curated content. IntAct is the default resource but all other PSICQUIC databases can be selected and in addition, users can submit custom interactions. Interaction data from IntAct are also included in the Reactome main search and the Analysis Service, helping users identify pathways of interest. In summary, Reactome has facilitated data integration by providing easy-to-use services and reusable widgets. Several resources such as OpenTargets (https://www.targetvalidation.org/), ChEBI (https://www.ebi.ac.uk/chebi/), BluePrint (http://dcc.blueprint-epigenome.eu/), PRIDE (http://www.ebi.ac.uk/pride/archive/), PINT (http://sealion.scripps.edu/pint/) and IP2 (http://goldfish.scripps.edu) have already integrated these services and widgets.

Indexing Clinical Research Datasets Using HL7 FHIR and Schema.org

Harold Solbrig, Mayo Clinic; *Guoqian Jiang, Mayo Clinic*

Schema.org was developed by a number of major search engines such as Bing, Google and Yahoo! as a common vocabulary for marking up web pages. The combination of HTML and Microdata, RDFa 1.1 Lite or JSON-LD enables a well-known set of semantic tags to be added to existing human-readable web pages. Schema.org has been widely adopted by public web sites and multiple extensions have been created for domains such as automobiles, bibliographic resources, product classifications, healthcare and life sciences.

The HL7 Fast Healthcare Interoperability Resources (FHIR) standard defines a standard set of "resources" that are used to exchange clinical and healthcare related information. FHIR is slated to become the de-facto inter-change mechanism for healthcare and related information. We have developed a schema.org representation for the FHIR information models known as fhir.schema.org. The purpose of this representation was to promote discussion of the value of fhir.schema.org to annotate web based clinical research datasets with their clinical model equivalent. We have initiated the collaboration with the Health and Lifesciences (http://health-lifesci.schema.org/) and Bioschemas.org (http://bioschemas.org/) groups to identify use cases and examine the relationships between these resources and healthcare data models.

Datasets2Tools: Enriching DataMed With Canned Analyses

Denis Torre, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Alexander Lachmann, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Biomedical data repositories, such as the bioCADDIE DataMed, enable the search and discovery of relevant research data digital objects. At the same time, tools that can operate on such data are indexed by repositories such as the Aztec.bio developed by the BD2KCCC. However, direct associations between datasets and tools are currently not available. Beyond such associations, it would be useful to systematically provide canned bioinformatics analyses for processed datasets on dataset landing pages. Here we present part of a pilot project to create a new type of digital object: canned analysis of a dataset with a specific online bioinformatics tool.

To enable the creation, management and usability of this new type of data object, we are developing the Datasets2Tools platform. Datasets2Tools includes a registry: a database of dataset-tool associations; a Google Chrome extension that displays icons that provide links to canned analyses from dataset landing pages of data repositories such as DataMed, the LINCS Data Portal (LDP) and the gene expression omnibus (GEO); and a web-site that enables users to register, browse, search, and grade dataset-tool associations. In the future, we hope that data repositories such as DataMed would adopt the Datasets2Tools protocol, making Datasets2Tools native to the DataMed site so users do not have to install the Chrome extension to view the icons for the canned analyses. By providing a simple and intuitive platform that links datasets to analysis tools, Datasets2Tools will lower the point of entry for many users of DataMed, LINCS LDP, GEO and other biomedical research repositories; the canned analyses will help these users extract more knowledge from such data repositories.

Aztec I: Building a Technology Platform to Integrate Biomedical Resources

***Justin Wood, UCLA**; Chelsea Ju, NIH BD2K Center of Excellence at UCLA and Department of Computer Science, UCLA; Patrick Tan, NIH BD2K Center of Excellence at UCLA and Department of Computer Science, UCLA; Brian Bleakley, NIH BD2K Center of Excellence at UCLA and Departments of Physiology, Medicine, and Bioinformatics, UCLA; Vincent Kyi, NIH BD2K Center of Excellence at UCLA and Departments of Physiology, Medicine, and Bioinformatics, UCLA; Tevfik Umut Dincer, NIH BD2K Center of Excellence at UCLA and Departments of Physiology, Medicine, and Bioinformatics, UCLA; Howard Choi, NIH BD2K Center of Excellence at UCLA and Departments of Physiology, Medicine, and Bioinformatics, UCLA; Peipei Ping, NIH BD2K Center of Excellence at UCLA and Departments of Physiology, Medicine, and Bioinformatics, UCLA; Alex Bui, NIH BD2K Center of Excellence at UCLA and Departments of Radiological Sciences, University of California, Los Angeles; Wei Wang, NIH BD2K Center of Excellence at UCLA and Department of Computer Science, UCLA*

Aztec, A to Z Technology, is designed to serve the BD2K awardees and to highlight their accomplishments in creating computational tools to facilitate knowledge translation. Aztec is a global biomedical resource that empowers users to simultaneously search a diverse array of resources including digital objects of databases, standalone software, web services, publications, and large libraries composed of many interrelated functions. In the current alpha version, Aztec offers over 5,430 tools and resources spanning domains as diverse as imaging, gene ontology, text-mining, data visualization and omics analyses.

To provide a reliable search platform, Aztec is hosted on an Amazon EC2 server, which will allow us to rapidly scale up to meet rising user demand. Aztec is implemented on a Node.js framework, using an enterprise established Model-View-Controller design pattern. Aztec employs a user management system with authentication and encryption to enable secure resource registration and updates. Aztec offers an advanced yet user-friendly query system that provides a cascade of nested sub-searches as well as comprehensive filtering and sorting of results. To preserve data consistency and integrity across various types of resources from diverse and voluminous inputs, Aztec employs a relational database implemented in MySQL. Powered by Solr from Apache Lucene, Aztec supports, in one unified platform, efficient keyword search and semantic query guided by established ontologies (see Aztec II). A professional interface with enterprise-strength visualization capabilities supported by Google Analytics provides superb user experience.

Future development of Aztec will leverage Neo4j to visualize rich graphs of tool relationships, such as reimplementations and workflows. In parallel, we will refine our weighting schema to provide better match to users' research needs. In summary, Aztec supports the BD2K Initiative awardees and biomedical community at large by ensuring the relevant resources remain findable and accessible, capturing all BD2K accomplishments in a sustainable manner.

Development of DataMed, a Data Discovery Index Prototype by bioCADDIE

Hua Xu, University of Texas Health Science Center at Houston; *Jeffrey S. Grethe, University of California, San Diego; Xiaoling Chen, University of Texas Health Science Center at Houston; Ruiling Liu, University of Texas Health Science Center at Houston; Ergin Soysal, University of Texas Health Science Center at Houston; Anupama Gururaj, University of Texas Health Science Center at Houston; Yueling Li, University of California, San Diego; Burak Ozyurt, University of California, San Diego; Hyeon-eui Kim, University of California, San Diego; Nansu Zhong, University of California, San Diego; Trevor Cohen, University of Texas Health Science Center at Houston; Todd Johnson, University of Texas Health Science Center at Houston; Mandana Salimi, University of Texas Health Science Center at Houston; Saeid Pournejati, University of Texas Health Science Center at Houston; Claudiu Farcas, University of California, San Diego; Alejandra Gonzalez-Beltran, University of Oxford; Philippe Rocca-Serra, University of Oxford; Cui Tao, University of Texas Health Science Center at Houston; Ian Fore, National Institutes of Health; Ronald Margolis, National Institutes of Health; George Alter, University of Michigan, Ann Arbor; Susanna-Assunta Sansone, University of Oxford; Lucila Ohno-Machado, University of California, San Diego*

Overview of the System

The primary objective of the DataMed team is to design and implement a DDI web application with a user-friendly interface that enables users to browse and search datasets that best satisfy their specific needs. DataMed consists of 3 main components:

- Repository ingestion and indexing pipeline that maps the disparate metadata from diverse data sources, into a unified specification provided by various working groups organized by bioCADDIE and related communities. This pipeline involves an automated component that provides controlled translation and curation of metadata using special tools such as transformation language and JSON-Path.
- Terminology server based on a graph database for terminological consistency that includes several specific terminologies and ontologies such as MeSH, Gene Ontology and NCBI Taxonomy, which are mapped via UMLS integration.
- Web application based on a search engine that uses the discovery index for locating appropriate datasets from the processed set of repositories. This engine also uses the terminology server to expand queries using synonymous terms.

The DataMed demonstration will cover an overview of the web interface to the search engine and will be based on a few common search scenarios. The aim is to introduce the system to the audience and to interact with and directly collect feedback from end users of the system.

Current Status: DataMed is currently available at http://datamed.biocaddie.org, encompassing 649,055 datasets from 23 different repositories that include 10 different data types. The CDT is working to incorporate new repositories to expand the content of the DDI as well as on validating, testing and improving ranking algorithms to retrieve relevant personalized search results. We are also conducting user and usability testing.

Acknowledgements. bioCADDIE has many more team members and collaborators than would fit the author list here. They are listed at http://biocaddie.org.

Metadata Mapping in bioCADDIE: Challenging Cases

Nansu Zong, Department of Biomedical Informatics, School of Medicine, University of California, San Diego; *Hyeon-eui Kim, Department of Biomedical Informatics, School of Medicine, University of California, San Diego*

The metadata mapping workflow in data ingestion pipeline of bioCADDIE (biomedical and healthcare Data Discovery Index Ecosystem) allows the different biomedical resources to be described with a unified and coherent metadata model, DATS (Data Tagging Suite). However, we encountered several challenges during the process due to the differences in how the metadata are structured and what information the metadata are designed to capture between various data repositories (i.e., source data repositories) and DATS. In this abstract we will introduce these challenges with specific examples as summarized below and discuss potential approaches to address them.

(1) Differences in granularities of metadata representation

Some metadata of source data repositories cannot be sufficiently represented with DATS thus potential information loss is expected after the translation. For example, the properties “country”, “latitude”, and “longitude” of the entity “geographic location” in American Gut Project are all mapped to the metadata item location in DATS v2.1.

(2) Implicit information on source repositories

The relational structure of DATS provides an efficient representation on complex information on a source repository such as the role of a dataset (i.e., “input” to an activity like data analysis, and “output” from the data analysis).

However, this information is only implicit in many source repositories. This would make the full automation of metadata extraction and mapping a bit challenging. For example, Human Microbiome Project (HMP) provides both raw (i.e., “input”) datasets and the processed (i.e., “output) datasets but without explicit flagging of these roles.

The challenges we faced consume a lot of human resource on metadata mapping and validation. They might also affect the search results in DataMed, the prototype data indexing and discovery tool of bioCADDIE. DATS model will continuously be refined and enriched based on these findings.

Software, Analysis, & Methods Development

Histology-Validated Neural Networks Enable Optical Coherence Tomography Virtual Histology

Vikram Baruah, University of Texas at Austin; *Aydin Zahedivash; Taylor B. Hoyt; Rogelio Salomon; Deborah Vela; L. Maximilian Buja; Thomas E. Milner; Marc D. Feldman*

Introduction

Although virtual histology algorithms using intravascular ultrasound data can successfully classify plaque composition, parallel results have not been demonstrated with intravascular optical coherence tomography (IVOCT). The complexity of IVOCT images has limited its clinical acceptance despite its greater resolution. For example, thin-capped fibroatheromas (TCFAs), which are uniquely recognized by IVOCT, can be falsely identified by expert viewers.

Hypothesis

We assessed the hypothesis that histology validated neural networks can reliably characterize arterial tissue and TCFA in IVOCT images.

Methods

Neural network features and nodes were optimized to best classify arterial fibrous, calcium and lipid in IVOCT images. Grey level co-occurrence matrix and windowed texture analysis was used to develop features. Lipid pixels were used to isolate TCFA. The network was trained on IVOCT pixels sampled from 21 plaques in 11 human hearts (3 women, 8 men) imaged within 24 hours of death. Accuracy was calculated by comparing network classification to histology assessment.

Results

Fibrous pixels were classified with 94.1% ± 0.38% sensitivity and 90.6% ± 0.24% specificity, calcium pixels with 87.5% ± 0.38% sensitivity and 91.8% ± 0.76% specificity, lipid pixels with 94.1% ± 0.39% sensitivity and 95.9% ± 0.24% specificity and TCFA pixels with 94.1% ± 0.52% sensitivity and 84.1% ± 0.32% specificity (n=170000). Accuracies reported herein exceed previously reported values for automated IVOCT plaque classification. Histology validation makes the presented method more reliable than expert observer-validated approaches.

Conclusion

We have developed a histology-validated IVOCT-based automated plaque classification algorithm that is able to colorize plaque composition in human coronary arteries. Gaining spatial insight into the composition of plaque can enable precise TCFA identification and provides a valuable diagnostic methodology to assess the efficacy of therapeutic interventions.

REproducible by Design, A Docker-Based Tool for Workflow Design, Data Linkage, and Workflow Execution

Tyler Bath, Department of Biomedical Informatics, University of California, San Diego; *Jihoon Kim, Department of Biomedical Informatics, University of California, San Diego; Lucila Ohno-Machado, Department of Biomedical Informatics, University of California, San Diego; Claudiu Farcas, Department of Biomedical Informatics, University of California, San Diego*

Despite calls for the repeatability and reproducibility in research, genomics community are not responding to these well, due to problems arising from obsolete source code, missing descriptions about the analyses, and unavailability of external resource datasets. Docker, a software containerization platform, was adopted to analyze thousands of cancer whole genome samples in International Cancer Genetics Consortium (ICGC). Docker encapsulated all the software components and guarantees to run the analysis software always the same in any computing environment, which enabled to run the identical workflow of alignment and variant calling in ICGC participating centers. Expanding this experience, we developed the REproducible by Design (RED) tool as a mechanism to assist in the analysis workflow design stage via Docker containers. RED retrieves the source code from remote repositories like GitHub, identify the set of programming languages used for development and automates the package building process. In case of successful builds, RED builds an appropriate Docker to encapsulate the relevant tools and all their dependencies. In case of failure, the existing tools for Continuous Integration are used to inform the author the encountered issues and possible solutions to make the build process less error-prone. A successfully built container is executed with example data through FlightDeck, a self-service web-portal that encapsulates scientific software as Virtual Machines (VMs) and enables automated VM provisioning based on published workflows and recipes. Working workflows and associated metadata about supported data are then stored into a secure cloud storage and assigned unique identifiers that the authors can use for publication. This process ensures that published science has an associated proven code and example data that can be reproduced regardless of later modifications to the original source code or data. RED was developed as BD2K supplement to the iDASH repository, supported by NIH/NHLBI National Centers for Biomedical Computing (U54HL108460).

The Georgetown Database of Cancer (G-DOC): A Web-Based Data Sharing Platform for Precision Medicine

Krithika Bhuvaneshwar, Georgetown University; *Anas Belouali, Innovation Center for Biomedical Informatics, Georgetown University; Shruti Rao, Innovation Center for Biomedical Informatics, Georgetown University; Adil Alaoui, Innovation Center for Biomedical Informatics, Georgetown University; Yuriy Gusev, Innovation Center for Biomedical Informatics, Georgetown University; Robert Clarke, Department of Oncology, Georgetown University; Louis M. Weiner, Department of Oncology, Georgetown University; Subha Madhavan, Innovation Center for Biomedical Informatics, Georgetown University*

An overarching goal of biomedical research is to improve the use and dissemination of rapidly growing biomedical datasets to support precision medicine. Individualized molecular profiling and the identification of predictive biomarkers can powerfully inform the choice of therapies for cancer patients. However, both require integration of extensive molecular, clinical, and pharmacological data, often from disparate and diverse sources. The Georgetown Database of Cancer (G-DOC) was designed and engineered to be a unique multi-omics data analysis platform to enable translational research and precision medicine.

G-DOC is home to 61 datasets that contain data from over 10,000 patients across 14 diseases (10 cancers and 4 non-cancers). 1700+ researchers from over 48 different countries worldwide currently use the platform. The data and tools in the G-DOC system have enabled over 40 research publications. G-DOC has the largest public collection of brain cancer patients from NCI Rembrandt dataset (671 patients).

G-DOC integrates clinical, transcriptomic, metabolomic, microRNA, next generation sequencing (NGS) data, and MRI medical images with systems-level analysis tools into a single, user-friendly platform. The “Variant Search” feature in G-DOC currently enables exploratory analysis of mutations based on genes, chromosomes, and functional location. A researcher can use this feature to 1) identify clinically actionable mutations in their dataset 2) identify pathways that may be affected by these mutations, and 3) identify novel mutations in their dataset and explore their potential impact on protein function.

We are currently working on developing features to support the import, integration, search, and retrieval of CLIA/CAP-certified cancer molecular diagnostic (molDx) data. This will enhance G-DOC’s interoperability with clinical and patient molecular profiling data that may be already stored in other databases. Our vision is to continuously improve and expand G-DOC with the long-term vision of supporting integration of informatics techniques into everyday research and practice.

Flexible Bootstrapping Approaches Toward the Clustering of Complex Medical Data

Rachael Blair, University at Buffalo*; Brian Chapman, University of Utah; Arianna DiFlorio, Institute of Psychological Medicine and Clinical Neurosciences; Ellen Eischen, University of Oregon; David Gotz, The University of North Carolina at Chapel Hill; Matthews Jacob, University of Iowa; Han Yu, State University of New York at Buffalo*

Identifying subgroups from a severely heterogeneous population is major challenge for Big Data. Different clustering methods optimize differently and consequently capture different aspects of relatedness in the population. Since there is not a one size fits all solution, and no gold standard, the selection of a clustering method can be daunting and problematic. Our interdisciplinary team is working towards the development of interactive ensemble methods for clustering Big Data.

In this first year, we have begun to lay the methodological foundation through the development of a non-parametric bootstrapping approach to estimate the stability of a clustering method. We have developed two novel approaches to bootstrapping stability, and accompanying visualizations, that accommodate different model assumptions, which can be motivated by an investigator's trust (or lack thereof) in the original data. Our approaches outperform state of the art methods for simulation and real data sets of moderate size.

A long-term vision of our work is to extend this bootstrapping approach to improve classification and diagnosis of mood disorders, in particular bipolar disorder and major depressive disorder, using data from the UK Biobank. This endeavor would require automated feature selection, sophisticated visualizations, and methods that accommodate mixed data, while retaining valuable clinical interpretations. This project is motivated by the hypothesis that a more precise and personalized classification of mental health disease can be obtained through the development of novel clustering methods that identify clinically significant structures with large population data sets.

KnowEnG: Cloud-Based Environment for Scalable Analyses of Genomic Signatures

Charles Blatti, University of Illinois at Urbana-Champaign; *Matthew Berry; Lisa Gatzke; Amin Emad; Nahil Sobh; Colleen Bushell; Saurabh Sinha*

Medical and biological researchers require advanced and convenient tools to cluster, characterize, and uncover important features from the genomic datasets they produce (e.g. transcriptomic, epigenetic, or genotypic profiling experiments). A body of recent research has shown these types of analysis are often improved by incorporating prior biological knowledge, such as known relationships and interactions between entities like genes, proteins, functional roles, disease phenotypes, etc. KnowEnG (the "Knowledge Engine for Genomics" strives to make the results of these advanced machine learning and graph processing algorithms, as well as dozens of collections of standardized biological knowledge datasets available to researchers through an intuitive user interface and interpretable visualizations. Our platform is designed to be deployed on heterogeneous, commercial cloud services and relies on technologies such as Docker and Apache Mesos to decompose large, complex analysis workflows and datasets into independent components to maximize parallel computation. In KnowEnG, multiple classes of analysis workflows for genomic datasets are available in both standard and knowledge-informed implementations including 1) gene prioritization methods for tasks like ranking genes for their role in drug response and 2) gene set characterization methods for finding relevant biological annotations. Through analysis available in our platform, we highlight novel insights gained from using random walks on prior knowledge networks over conventional approaches.

Interactive Web Application for Visualization of Brain Connectivity

***David Caldwell, University of Washington**; Jing Wu; Kaitlyn Casimo; Jeffrey Ojemann; Rajesh P.N. Rao*

We present here a novel, lightweight, web-based application for visualizing patterns of connectivity between 3D stacked data matrices with large numbers of pairwise relations. Visualizing a connectivity matrix, looking for trends and patterns of interest, and subsequently dynamically manipulating these values is a challenge for scientists from diverse fields, including neuroscience and genomics. Neuroscience data sets with high-dimensional connectivity data include those acquired via EEG, electrocorticography, magnetoencephalography, and fMRI. We demonstrate the analysis of connectivity data acquired via human electrocorticography recordings as a domain-specific implementation of our application. Neural connectivity data often exists in a high-dimensional space, with multivariate attributes for each edge between different brain regions, which motivates a lightweight, open-source, easy to use visualization tool to allow for the rapid exploration of these connectivity matrices to highlight connections of interest. Here we present a client-side, mobile-compatible visualization tool written entirely in HTML/CSS/JS that allows for in-browser manipulation of user-defined files for the exploration of brain connectivity. Visualizations can highlight different aspects of the data simultaneously across differing dimensions, allowing for rapid exploration and analysis of the data. Input files are in JSON format, and custom Python scripts have been written to allow for the parsing of MATLAB or Python data files into JSON-loadable format. We envision applications for this interactive tool in fields ranging from neuroscience to genomics to any other field seeking to visualize pairwise connectivity.

GRcalculator: An Online Tool for Calculating and Mining Drug Response Data

Nicholas Clark, University of Cincinnati; *Marc Hafner, Harvard Medical School; Michal Kouril, University of Cincinnati; Mario Niepel, Harvard Medical School; Elizabeth Williams, Harvard Medical School; Peter Sorger, Harvard Medical School; Mario Medvedovic, University of Cincinnati*

Large-scale dose-response data and genomic datasets can be combined to discover novel drug-response biomarkers. There exist numerous online datasets of drug-response assays, but they are currently poorly accessible and their potential as a big data resource is limited due to lack of access and connection. Furthermore, it has recently been found that drug-response data often vary from one study to the next. A major reason for this variance is that traditional metrics of drug sensitivity such as IC50, Emax, and AUC values are confounded by the number of cell divisions taking place over the course of an assay. To solve this problem, we have developed GRcalculator, a suite of online tools found at www.grcalculator.org. The tools use GR metrics (proposed recently by Hafner et al. in Nature), a set of alterative drug-response metrics based on growth rate inhibition that are robust to differences in nominal division rate and assay duration.

GRcalculator is a powerful, user-friendly and free tool for mining drug-response data using GR metrics. These metrics harmonize drug-response data, improving the discovery of novel drug-response biomarkers using big data as well as allowing for comparisons with patient-derived tumor cells that are generally slow growing. Direct access to LINCS drug-response datasets and, in the future, other public domain datasets is a unique functionality that will facilitate re-use of the valuable resources that these data represent. As well as mining datasets, the tool also offers calculation and visualization of GR metrics (and traditional metrics), generates publication-ready figures, and provides a unified platform for researchers analyzing drug sensitivity. For offline calculation and analysis, we have developed the GRmetrics R package (available via Bioconductor), which allows for use of larger datasets and inclusion of GR metrics calculations within existing R analysis pipelines.

Formal Evidence Networks for Reproducibility in Biomedical Translation

Tim Clark, Massachusetts General Hospital and Harvard Medical School

Therapeutic targets growing out of academic research have a dismal record of translation into pharmaceutical drug discovery and development. Attrition on transfer to pharma is reported to range from 75% to 89%, followed by further attrition up to an additional 96% once in the pharmaceutical pipeline. This lack of robustness is clearly a waste of scarce research resources and has led to serious concern by NIH.

Robust validation of proposed biological targets requires analysis of dozens to scores of detailed hypotheses. These in turn may be based (ultimately) on findings from up to thousands of academic research publications. The many layers of information involved constitute deep networks of hypotheses, assertions, and evidence that at present are poorly represented and extremely laborious to analyze with sufficient rigor.

If the underlying network of evidence is insufficiently understood, drug targets may be advanced at great cost on insufficient or flawed evidence.

Formal argument graphs, or claim-evidence networks, have an extensive literature in Artificial Intelligence. We describe how these models may be applied to creating computable summaries of publication corpora; to finding and highlighting incomplete or flawed evidence chains across publications; and to building formal target validation models across the (perilous) translational “gap” between academia and industry.

We also show how implementing data and resource citation and indexing at scale can create, for any publication, the core of a formal argument graph.

Augmenting Metadata With Models of Experimental Methods

Scott Colby, Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University; *Mark Musen, Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University*

A rigorous and machine-readable model for representing experimental methods is a necessity for improving the replicability and reproducibility of scientific experiments. Knowledge about experimental methods can be incorporated as metadata in online data sets and used to guide the work of future investigators. The technology being developed by the Center for Expanded Data Annotation and Retrieval (CEDAR), which is designed to ease the annotation of online data sets, will use such representations of experimental methods to improve communication among researchers and to automate experimental workflows that replicate previous investigations. We have identified three distinct levels of abstraction used to represent experimental methods: (1) the level of the "abstract" (how the method would be presented in a paper's abstract), (2) the level of the "methods section," and (3) the level of the "notebook" (or that of supplementary information). Even at the level of "supplementary information," important details needed to carry out the experiment are often left out because they are assumed to be obvious. This problem is only exacerbated at the less-specific levels. With a rigorous model of this experiment, all steps, no matter how "well-known," will be expressed. Having all details available, if needed, will allow for better understanding of the method and for better understanding of the resultant data. Our work will result in the identification of motifs and steps in protocols as they are written at all three levels. The representation scientific methods in this manner will provide a detailed, multi-scale model of scientific procedures that can stand in for the methods section of a publication when searching for or comparing data sets online.

Fast, Accurate Causal Search Algorithms From the Center for Causal Discovery (CCD)

Gregory Cooper, University of Pittsburgh; *David Danks, Carnegie Mellon University; Joseph Ramsey, Carnegie Mellon University; Peter Spirtes, Carnegie Mellon University; Clark Glymour, Carnegie Mellon University*

Computational procedures for extracting from data causal relationships represented as directed graphs in a variety of scenarios (non-experimental data; no confounding; unknown confounding; sample selection bias; feedback) have been available for very low dimensional problems for two decades. In the last two years the CCD has, among other things, redesigned, parallelized, combined, and re-implemented these procedures so that they produce accurate results for problems with very high dimensions (up to 106 variables) and low sample sizes (103).

Sparse, Gaussian models without latent variables or cycles and with a million variables can now be recovered using the Fast Greedy Search (FGS) algorithm (a redesign and parallelization of the Greedy Equivalence Search method) to datasets with 1,000,000 variables with better than 98% precision in less than 18 hours using Pittsburgh Supercomputing Center resources. 50,000 variable problems can be solved with comparable precision on a 4-core laptop in less than 15 minutes. These procedures have been used on voxel level fMRI data, which are presumably very dense; to show that enforcing sparsity does not reduce precision. Another innovation, PC-Max provides similar improvements in speed and accuracy for PC, the oldest correct search algorithm for directed acyclic graphs.

The Fast Causal Inference (FCI) algorithm and its variants recover causal relations among measured variables when there may be latent confounders of the measured variables. Using FGS as a preprocessor, FCI has been speeded up and its causal-discovery accuracy much improved.

Current work includes developing and testing improvements of the Cyclic Causal Discovery algorithm, generalizations of FGS and PC-Max to non-Gaussian, non-linear systems, SAT solver search algorithms, under-sampled time series, mixed data types, and a Bayesian algorithm for finding expression pathways.

Old Medicines, New Uses: Upcycling Drugs Using Social Media and Cheminformatics

Nabarun Dasgupta, Epidemico, Booz Allen Hamilton*; Charles Schmitt, The University of North Carolina; Weifan Zheng, North Carolina Central University; Eugene Muratov, The University of North Carolina; Alexander Tropsha, The University of North Carolina*

Drug repurposing is the practice of finding new uses for existing medications through clinical observation and chemical modeling. In medical fields such as pediatric oncology, discovering off-label benefits can rapidly and efficiently reveal new uses for drugs that may otherwise not be viable targets of commercial exploration by drug manufacturers. However, patient reported benefits from off-label use are notoriously difficult to compile on a large scale, especially since patients may not report them to physicians. The massive amount of public conversations on social media can be used to create a novel database representing patient experience with off-label use. Public social media posts mentioning drug names (and misspellings) can be purchased or scraped. Most importantly, internet vernacular must be translated to standardized medical terminology using natural language processing (NLP), and relevant posts isolated using machine learning (ML). While semantic approaches have shown promise, the NLP used in this project relies on manually collected dictionaries (English, Spanish, French) where tens of thousands of colloquial words, phrases and emoji have been collected and linked to medical terminology ("I look like a beet" would translate to erythema). The Bayesian ML process relies on a database of a half-million social media posts that have been manually curated to select those that contain side effect information. The platform has been successfully demonstrated to identify previously unknown adverse events for medicines, and is in use for routine surveillance of portfolios of drugs by major pharmaceutical companies. In this project we extend social listening to drug repurposing using a chemcocentric approach, whereby social media insights are connected to chemical structures and then input into Quantitative Structure-Activity Relationship (QSAR) models. This approach blends advanced chemical modeling and text mining to efficiently identify new drug targets for the some of the most challenging clinical shortcomings in the modern pharmacopeia.

A Software Suite for Causal Modeling and Discovery

Jeremy Espino, University of Pittsburgh; *Joseph Ramsey, Department of Philosophy, Carnegie Mellon University; Kevin Bui, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Chirayu Wongchokprasitti, Department of Biomedical Informatics, School of Medicine/School of Information Sciences, University of Pittsburgh; Zhou Yuan, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Mark Silvis, Department of Computer Science, University of Pittsburgh; Michael Davis, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Jitske Venema Shunfenthal, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Nick Nystrom, Pittsburgh Supercomputing Center/Carnegie Mellon University; Alexandros Labrinidis, Department of Computer Science, University of Pittsburgh; Panos K. Chrysanthis, Department of Computer Science, University of Pittsburgh; Gregory Cooper, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh*

In the past 25 years, tremendous progress has been made in developing general computational methods for discovering causal knowledge from data based on a representation called causal Bayesian networks. While much progress has been made in the development of these computational methods, they have not been readily available, sufficiently efficient, or easy to use by biomedical scientists, and they have not been designed to exploit Big Data that are increasingly available for analysis.

The Center for Causal Discovery has created a suite of tools that make efficient causal modeling and discovery (CMD) algorithms from Big Data available on a variety of platforms and environments. The suite uses a common set of CMD algorithms implemented as a Java library. We have created software around this library to develop our suite:

- Tetrad-lib – a readily imported Java library of CMD algorithms
- Tetrad – a desktop application that runs on any Java-enabled computer
- Causal-cmd – a command-line application that runs on any Java-enabled computer
- Causal-web – an easy-to-use Web-based application that submits causal discovery jobs to an HPC (e.g., three terabyte 64 core nodes at the Pittsburgh Supercomputing Center (PSC) or Amazon EC2)
- R-causal – an R library
- Py-causal – a Python module
- A Docker instance that contains a ready-to-run instance of R-Studio and the R-causal
- Causal-REST-API – a RESTful API hosted at 1) the PSC that submits CMD jobs to the Bridges supercomputer and 2) Amazon EC2

All of our software is open source and licensed under the GNU GPL such that it can be modified and incorporated into other software. The software and documentation is freely available from www.ccd.pitt.edu, and the source code is available from github.com/cmu-phil/tetrad and github.com/bd2kccd.

Clustergrammer: Interactive Visualization and Analysis Tool for High-Dimensional Biological Data

Nicolas Fernandez, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Hierarchically clustered heatmaps are a popular visualization technique to display high-dimensional biological datasets as figures in research publications. To create such graphics, experimental biologists often rely on computationally oriented collaborators. Moreover, most tools developed to visualize hierarchically clustered heatmaps generate static images. Clustergrammer is a web-based matrix visualization tool, built using the D3 JavaScript library to enable experimental and computational-biologists to easily generate highly interactive heatmaps. Clustergrammer implements many interactive features including: zooming, panning, filtering, sorting, reordering, dimensionality reduction, gene set enrichment analysis, gene description text, and shareable persistent links. At the meeting, we will demonstrate how to use Clustergrammer to visualize and analyze a diverse set of high-dimensional biological data including: gene expression, protein-protein interactions, and CyTOF single-cell proteomics data. Clustergrammer provides interactive visualizations for several BD2K-LINCS DCIC web-tools including: Enrichr, GEN3VA, GEO2Enrichr, Harmonizome, and L1000CDS2. In addition to the web app, Clustergrammer can be embedded with IPython Jupyter Notebook or as a standalone open-source JavaScript or Python library. A tutorial on how to use Clustergrammer is provided at: http://amp.pharm.mssm.edu/clustergrammer/

Weak Supervision: Biomedical Entity Extraction Without Labeled Data

Jason Fries, Stanford University; *Sen Wu, Stanford University; Alex Ratner, Stanford University; Christopher Re, Stanford University*

Recent advances in representation learning are largely replacing manual feature engineering in natural language processing systems. Since current deep learning approaches require large, labeled datasets to learn features, training set creation is now a key bottleneck in building information extraction applications. For many tasks, however, there is a wide spectrum of structured resources like formal ontologies and domain expertise that can be leveraged as forms of weak supervision to heuristically generate large-scale datasets with noisy labels, which are then used to train deep learning systems. In this work we built a named entity recognition (NER) system for tagging disease and chemical names in PubMed abstracts. Unlike traditional NER approaches, we do not use manually labeled training data, instead relying on sets of heuristic labeling functions to programmatically create weak labels. We build upon recent work formalizing a generalization of distant supervision called data programming [1] which allows us to learn over collections of noisy and conflicting heuristic rules to train a generative model of the underlying annotation process. This model allows us to programmatically construct “denoised” training data over large collections of unlabeled data, which are then used to train downstream discriminative models using logistic regression and deep learning approaches. Using this method on three public biomedical datasets, we find that it is possible to approach benchmark supervised learning performance without using hand-labeled data, instead relying on weak supervision provided by ontologies and other heuristics. This suggests that data programming is a potentially viable way of building extraction systems for the long tail of biomedical concept types for which there is little-to-no traditionally labeled training data.

[1] Ratner, Alexander, et al. ""Data Programming: Creating Large Training Sets, Quickly."" arXiv preprint arXiv:1605.07723 (2016).

Reproducible Exploratory Data Analysis With Vistories

Nils Gehlenborg, Harvard Medical School; *Alexander Lex, University of Utah; Samuel Gratzl, Johannes Kepler University Linz; Marc Streit, Johannes Kepler University Linz*

Visualization is a key data analysis approach that allows scientists to explore datasets without preconceived questions, and is thus crucial for hypothesis generation. When combined with algorithmic approaches, it bridges the gap between exploration and confirmation. Visualization is also essential in communicating research findings.

Current visualization tools, however, have a crucial shortcoming: the interactive visual exploration process is not captured, which means that the analysis steps cannot be shared. Being able to reproduce visual analysis sessions and enabling third parties to understand, modify, and extend analysis sessions can have a significant impact on transparency, reproducibility, and innovation of analysis processes. Furthermore, there is enormous potential to utilize visual analysis sessions to efficiently communicate data.

We developed methods and tools that make this vision a reality. By capturing the visual analysis process and by enabling users to comment on their decisions, we make visual analysis reproducible and sharable. We also leverage data about the analysis process to allow scientists to create ""Vistories"", which are interactive and annotated figures, to communicate their findings. Vistories do not only efficiently communicate the findings, but also give audiences the ability to re-trace and modify an analysis.

We demonstrate the Vistories approach as an extension of the StratomeX cancer subtype analysis tool to illustrate how clustering and other stratification techniques can be enhanced by capturing multiple possible solutions. Demos for our prototype for this approach are accessible at http://vistories.org.

TruenoDB: A Network Database System for Managing, Analyzing, and Querying Large Biological Networks

Ananth Grama, Purdue University; *Victor Santos, Purdue University; Servio Palacios, Purdue University; Edgardo Barsallo, Purdue University; Miguel Rivera, Purdue University; Peng Hao, Purdue University; Chih-Hao, Purdue University; Wojciech Szpankowski, Purdue University; Tyler Cowman, Case Western Reserve University; Mustafa Coskun, Case Western Reserve University; Joseph Ledger, Case Western Reserve University; Zachary Stanfield, Case Western Reserve University; Nirupama Bhattachary, University of California, San Diego; Shamim Molla, University of California, San Diego; Sindhushree Raghunandan, University of California, San Diego; Shankar Subramaniam, University of California, San Diego; Mehmet Koyuturk, Case Western Reserve University*

Ever-increasing amounts of physical, functional, and statistical interaction data among bio-molecules, ranging from DNA regulatory regions, functional RNAs, proteins, metabolites, and lipids, offer unprecedented opportunities for computational discovery and for constructing a unified systems view of the cellular machinery. These data and associated formalisms have enabled systems approaches that led to unique advances in biomedical sciences. However, storage schemes, data structures, representations, and query mechanisms for network data are considerably more complex, compared to other, "flat" or low-dimensional data representations (e.g., sequences or molecular expression). For this reason, the query and application programming interfaces (API) offered by existing biological network databases are limited to descriptive information on the connections in the network. There exist commercial database systems that are dedicated graph data (e.g., Neo4j, TitanDB); however, these general-purpose systems lack necessary support for biological network databases, do not have significant support for statistical modeling, have minimal analytics capabilities, and are not extensible/ customizable. Here, we describe TruenoDB, a distributed graph database system that is developed specifically for integrated querying and analysis of biological network data.

TruenoDB's storage system is built on Apache Cassandra and it implements a computing engine integrated with built-in network analysis and exploration algorithms. The high performance computing engine implemented in TruenoDB is based on Spark/GraphX and it provides online transaction processing (OLTP) support for Gremlin traversal language. We demonstrate the API and web-based query interface of TruenoDB in the context of processing network proximity queries on multiple versions of the Intact Protein Interaction Database.

Increasing NCBO BioPortal and CEDAR Synergy for BD2K

John Graybeal, Stanford University; *Jennifer Vendetti, Stanford University; Michael Dorf, Stanford University; Martin O'Connor, Stanford University; Marcos Martínez-Romero, Stanford University; Mark Musen, Stanford University*

The NCBO BioPortal provides Web services for over 500 biomedical ontologies, allowing investigators to annotate and retrieve data, generate value sets, and perform advanced analytics of a wide range of biomedical and clinical data. BioPortal provides core services for many CEDAR metadata activities, and also heavily serves the wider BD2K community—seven out of eight centers responding to the BD2K metadata survey cited BioPortal as an essential resource for their work.

We have completed some long-planned features enabling or optimizing CEDAR's use of BioPortal. With this year's supplement, we have begun pursuing enhancements to maximize BioPortal's value to CEDAR and advance BioPortal's integration with other BD2K programs and services. We present the work accomplished, and enhancements that are being made available over the course of this supplement.

The first set of completed features advanced CEDAR's usage of BioPortal concepts and value sets. We made subtle improvements to BioPortal's ontology presentation services and added features for accessing and extending value set services.

We also began enhancing BioPortal's term-specific services for use by CEDAR and others. We have prototyped a concept-centric view of ontology data, providing for a given term all significant information available across all BioPortal ontologies as well as other relevant sources. Using this baseline, we plan to provide a "best-term" identification service, taking into account more contextual information specific to term discovery; and a query-expansion service.

As CEDAR adds property relations for its fields and terms, we will advance BioPortal's handling of ontology properties (the formal relationships between terms), extending the BioPortal API to better handle and retrieve ontology properties. We'll also be increasing BioPortal's metadata-handling capabilities.

Finally, we intend to harden BioPortal's capacity for complex queries by CEDAR, both by optimizing query handling (partially completed), and developing application-level test suites around CEDAR's most important queries.

Establishing Context: Geospatial Health Context Cube

Timothy Haithcoat, University of Missouri, MU Informatics Institute

When developing a research approach or developing a clinical trial, it is important that the selection of areas, cases or subjects for the research be understood to the highest degree possible. The developed data structure assembled within a raster data representation creates a new tool through which the context for a study can be examined and analyzed. The data comprises over 300 layers that have been captured from existing public data sources and re-represented through interpolation, aggregation, and generalization into a continuous surface at a given spatial resolution. These data include demographic, social, economic, educational, cultural, infrastructural, and environmental information. At the same time, these data have also been mapped into two parallel spatial structures, but rather than the attribute value itself being mapped, its uncertainty has been mapped in two dimensions: the spatial dimension (X, Y) and the attributes own margin of error. By using this parallel structure, any selection, modeling, or analysis conducted can produce two related products: the result itself, and the confidence of that result based on both attribute and spatial uncertainties. This is accomplished conducting a simple intersection of the health data points (or areas) with the Health Context Cube. This simple spatial intersection will calculate and add attributes from all of these layers of information to a developing sampling scheme. Statistical analysis can then be performed and the contextual or spatial relatedness among the samples can be assessed. The ability to 'know' and possibly choose to control for 'outside' variables (such as environmental, social, cultural, infrastructure, or other factors) during the development of a study or trial can provide a clearer or cleaner picture of the health aspect under investigation. We believe that this will lead to higher correlations and stronger research results in the health field.

Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification

Sanda Harabagiu, University of Texas at Dallas; *Travis Goodwin, University of Texas at Dallas; Ramon Maldonado, University of Texas at Dallas; Stuart Taylor, University of Texas at Dallas*

The annotation of a large corpus of Electroencephalography (EEG) reports is a crucial step in the development of an EEG-specific patient cohort retrieval system. The annotation of multiple types of EEG-specific clinical concepts is challenging, especially when automatically performed on Big Data. To address this challenge, we developed a novel framework which combines the advantages of active and deep learning while producing annotations that capture a variety of attributes of clinical concepts. The automatic annotation of the big data of EEG reports was performed by a Multi-task Active Deep Learning (MTADL) paradigm aiming to perform concurrently multiple annotation tasks, corresponding to the identification of (1) EEG activities and their attributes, (2) EEG events, (3) medical problems, (4) medical treatments and (5) medical tests mentioned in the narratives of the reports, along with their inferred forms of modality and polarity.

An important step of the MTADL paradigm was the design of the deep learning architectures capable to identify EEG-specific clinical concepts. We experimented with two deep learning architectures. The first architecture aims to identify (1) the anchors of all EEG activities mentioned in an EEG report; as well as (2) the boundaries of all mentions of EEG events, medical problems, medical treatments and medical tests. The second architecture was designed to recognize (i) multiple attributes considered for each EEG activity, as well as (ii) the type of the EEG-specific medical concepts. In addition, the second deep learning architecture identifies the modality and the polarity of EEG-specific concepts. The ability to learn jointly multiple types of concepts and attributes was made possible by the sampling mechanism used in the MTADL paradigm, based on the rank combination protocol, which combines several single-task active learning selection decisions into one.

BDDS Tools to Enable Transcriptional Regulatory Network Analysis

Ben Heavner, Institute for Systems Biology, BDDS Team

Transcriptional Regulatory Network Analysis (TReNA) is a functional genomics approach being developed by the Nathan Price Lab at the Institute for Systems Biology. This method presents several big data challenges that are being addressed by the Big Data to Discovery Science (BDDS) consortium: BDDS has built generalizable tools for the bulk transfer and reproducible processing of DNase-seq data made available by the Encyclopedia of DNA Elements data coordinating center (ENCODE DCC).

Data is selected through the ENCODE search interface, then assembled into a "Big Data Bag" with an ENCODE to BD Bag webservice (http://encode.bdbag.org/). BD Bags, implemented by BDDS, facilitate verifiable data assembly and transfer. Tools for using and working with BD Bags are publicly available (http://bd2k.ini.usc.edu/tools/bdbag/). The webservice identifies BD Bags of ENCODE data using the Minimal Viable Identifier (MINID), also developed by BDDS and publicly available at http://bd2k.ini.usc.edu/tools/minid/. This approach of assembling data into a BD Bag and uniquely identifying the data set with a MINID is generalizable to any collection of digital objects, not just data from ENCODE.

A researcher can use a MINID identifying a BD Bag of ENCODE data (or DNase-seq data from any source) to instantiate a data processing workflow implemented on the BDDS Galaxy Platform (https://bdds.globusgenomics.org/) to conduct processing using scalable Amazon Cloud resources. Beginning with the MINID, the Galaxy tool will retrieve the data directly from the ENCODE portal, obviating the need for cumbersome data transfers to local resources. From the researcher's perspective, one only need a MINID to begin the process of genomic alignment, peak finding, and footprinting with a variety of bioinformatic tools. Each instantiation of the workflow is identified by a MINID, and results are available via BD Bags.

Biobank to Digibank: High-Frequency Mobile Sensor Data Collection for Long-Lasting Research Utility

***Timothy Hnat, University of Memphis**; Syed Monowar Hossain, University of Memphis; Karen Hovsepian, Troy University; Nazir Saleheen, University of Memphis; Hillol Sarker, University of Memphis; Tyson Condie, University of California, Los Angeles; Emre Ertin, Ohio State University; Jim Rehg, Georgia Institute of Technology; Ida Sim, University of California, San Francisco; Mani Srivastava, University of California, Los Angeles; Santosh Kumar, University of Memphis*

For long-lasting research utility, biomedical studies often archive biospecimens in biobanks so that they can be reprocessed to take advantage of future improvements in assays and support biomedical discoveries not possible at the time of data collection. mHealth studies, on the other hand, usually collect digital biomarkers (e.g., activity counts) that are specific to the computational models adopted by respective vendors at the time of data collection. This approach prevents any future validation of these biomarkers and makes it impossible to recompute newer biomarkers. To obtain similar long-lasting research utility as biobanks, raw sensor data must be collected that can be reprocessed in future to validate prior biomarkers and to obtain newer biomarkers. Doing so is, however, challenging due to high frequency, large volume, rapid variability, and battery life limitations.

MD2K Center of Excellence has successfully developed open-source software (for both mobile phones and the cloud) that allow collection of high-frequency raw sensor data. The smartphone software called mCerebrum supports concurrent collection of streaming data from 8+ wearable sensors including: Microsoft Band, MotionSense, EasySense, AutoSense, Phone sensors (e.g., GPS), Omron Weight and Blood Pressure, and Oral-B smart toothbrush. It supports high-frequency raw sensor data collection (at 800+ Hz for 70+ million samples/day), curation, analytics, storage (~2GB/day), and secure uploads to a cloud. mCerebrum continuously assesses data quality to quickly detect and fix any data quality issues to correct sensor detachment or sensor misplacements on the body. Data science research conducted by MD2K has resulted into 10 mHealth biomarkers - stress, smoking, craving, eating, lung congestion, heart motion, location, activity, driving, and drug use. Several of these biomarkers are computed in real-time on the phone to support biomarker-triggered Just-in-Time Adaptive Interventions (JITAI). The demo will showcase live data collection and processing with MD2K sensors and software.

The Duke Data Service: Building an Infrastructure for Data and Provenance Microservices

Erich Huang, Duke University School of Medicine; *James Fayson; Matthew Gardner; Darin London; Darrin Mann; Donald Murry; Jonathan Parrish; Jonathan Turner; Nancy Walden*

Traditional academic health systems and medicine schools are highly silo-ed. Functions for research data storage and management are often scattered across multiple platforms and have differing security and format requirements. In order to bring World Wide Web Consortium (W3C) Provenance into life sciences research, and in order to make the relevant data and provenance services interchangeable across the many domains of a biomedical research institution, we are building a ""Data Service"" that is architecturally based on the concept of microservices that can be consumed in whole or in part by ""service clients"". One such client is a web client that employs Facebook's React framework for providing a user-friendly interface to Data Service functions. Another client is a command line Python client being built independently by the Duke University Center for Genomic & Computational Biology (GCB). The GCB Python client is a good example of how making the Data Service microservice APIs available to the public creates an ecosystem for developing service clients adapted to specific use cases.

The Data Service currently provides an OpenStack Swift object store hosted by Duke's Office of Information Technology. Currently this storage "endpoint" is in a non-PHI zone. A PHI-safe instance will be instantiated and connected in the next few months. The Data Service's CRUD (Create Read Update Delete), Provenance, Metadata, Neo4J, Postgres, and ElasticSearch business logic are maintained on the Heroku Platform-as-a-Service, and Shibboleth authentication is managed within Duke.

Currently in Alpha, the Data Service's initial large-scale deployment will occur in Duke's Center for Genomic and Computational Biology (GCB) and the Duke Human Vaccine Institute (DHVI). The Data Service will serve as the "delivery vehicle" for the GCB Sequencing and Genomic Technology and the DHVI Flow Cytometry Cores.

The Data Service is Open Source hosted at (https://github.com/Duke-Translational-Bioinformatics/duke-data-service) under GNU GPL3 license.

A Computational Framework for Identifying New Treatment Options in Glioblastoma

Haruka Itakura, Stanford University

Glioblastoma (GBM) is the most common and highly lethal primary malignant brain tumor in adults. There is a dire need for both effective therapeutic options and methods to identify patients who are most likely to respond to such therapies. GBM overexpresses vascular endothelial growth factor A (VEGF-A) and early phase clinical studies have supported the use of its inhibitor, bevacizumab. However, adding bevacizumab to standard GBM therapy did not improve survival in two recent phase 3 trials. In our study, we hypothesize that a methodological framework using quantitative magnetic resonance (MR) imaging features could identify within one of the clinical trial cohorts a subgroup of patients who possess the highest molecular susceptibility to bevacizumab therapy and will have experienced a survival benefit. We previously reported on a computational methodological framework that identified three novel GBM subtypes on the basis of quantitative MR imaging features, where each subtype also exhibited distinct molecular activity profiles that are potentially targetable as individualized therapies. The framework applied in two independent patient cohorts (single-institution and The Cancer Genome Atlas) identified a unique set of GBM imaging features that were linked to VEGFR pathway uprepregulation. We will apply the framework on MR images from the recent clinical trial patient cohort to determine whether patients with and without these imaging characteristics experienced a greater benefit in survival from bevacizumab use. The validation of our hypothesis would potentially lead to a new therapeutic option for a subgroup of GBM patients, as well as support the robustness of our methodological framework for broader use of in other cancers with imaging and molecular data.

SigNetA: Web Tool for Network Analysis of Gene Expression Signatures

Rashid Karim, University of Cincinnati; *Mario Medvedovic, University of Cincinnati*

Network analysis of biological systems is increasingly gaining acceptance as a useful method for integration and analysis of gene expression data. SigNetA (Signature Network Analysis) is a web application which provides a biologist-friendly user interface for interactive network analysis of large gene expression signature datasets. Users can upload gene IDs, p-values and fold change differential expressions of a gene expression signature to create an optimal subnetwork from Interactome or STRING human proteome datasets. The user is able to view the subnetwork, download generated network data, perform and download GO (Gene Ontology) enrichment analysis. All these functionalities have been integrated within SigNetA to facilitate the interpretation of the subnetwork.

Building Entity Matching Management Systems for Data Science Problems in Biomedicine

Pradap Konda, University of Wisconsin-Madison; *Sanjib Das, Department of Computer Sciences, University of Wisconsin-Madison; Paul Suganthan G.C., Department of Computer Sciences, University of Wisconsin-Madison; AnHai Doan, Department of Computer Sciences, University of Wisconsin-Madison; Adel Ardalan, Department of Computer Sciences, University of Wisconsin-Madison; Jeffrey R. Ballard, Department of Computer Sciences, University of Wisconsin-Madison; Han Li, Department of Computer Sciences, University of Wisconsin-Madison; Haojun Zhang, Department of Computer Sciences, University of Wisconsin-Madison; David Page, Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison; Eric LaRose, Marshfield Clinic Research Foundation; Jonathan Badger, Marshfield Clinic Research Foundation; Peggy Peissig, Marshfield Clinic Research Foundation*

Entity matching finds data that refer to the same real-world entities, such as (David Smith, John Hopkins University) and (Dave M. Smith, JHU). It is a pervasive problem in data science in general and in biomedicine in particular. Numerous entity-matching solutions have been proposed. These solutions however have focused mostly on the matching step, that is, developing algorithms that match accurately and fast.

In practice, entity matching is often a long and iterative process that involves many more steps, such as understanding and visualizing the data to be matched, developing, applying, and debugging blocking and matching, and understanding and visualizing the matching results, among others. There has been very little work on building tools to support the entire entity matching process. This work describes Magellan, an entity matching management system that supports the user in the entire end-to-end matching process. We describe how Magellan has been applied to match drug reference tables at Marshfield Clinic Research Foundation. The system found 4.3M matches between two tables of 453K and 451K tuples, respectively, achieving a precision of 99.18% and recall of 95.29%. Magellan has been released as an open source software and is available for the biomedical community to use.

Dashboard Visualization and Tool Integration for Enrichr

Maxim Kuleshov, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Genomic, transcriptomic, epigenomic, proteomic and other omic studies generate lists of genes and proteins that are hard to interpret. One powerful approach to help with the understanding the biological functions embedded within such lists is to query these lists against annotated gene sets for enrichment analysis. The BD2K-LINCS DCIC is actively developing and maintaining a popular gene set enrichment analysis tool called Enrichr. Since it publication, Enrichr was accessed by over 45,000 users who uploaded over 1.7 million lists for analysis with this tool. In 2016 the submission rate increased substantially, and it is currently at a median of more than 850 lists per day. Here we describe a new alternative method to visualize enrichment results with Enrichr, as well as integration with other tools. We plan to demonstrate how the results from the enrichment analyses can be visualized as a dashboard that displays summarized results for each biological category. Users can customize their dashboard by selecting the gene set libraries that they deem most relevant to their work. To infer upstream regulatory networks from differentially expressed genes, Enrichr will be integrating analysis from three related tools we have developed: Expression2Kinases (X2K), Genes2Networks (G2N) and L1000CDS2. X2K produces inferred upstream regulatory networks made of transcription factors, intermediate proteins, and protein kinases which are predicted to regulate lists of differentially expressed genes. G2N creates subnetworks from input genes or proteins based on known protein-protein interactions. L1000CDS2 produces lists of small molecules that can either reverse or mimic gene expression signatures based on the LINCS L1000 data. Overall, the integration of these tools into Enrichr, armed with a new dashboard visualization, is expected to enhance users experience to help investigators extract more knowledge from their data. Enrichr is freely available at http://amp.pharm.mssm.edu/Enrichr.

Gene Wiki Knowledgebase and Tool Development for Molecular Signatures of Cardiovascular Phenotypes

Jessica Lee, Heart BD2K Center at UCLA; *Sarah B. Scruggs, Heart BD2K Center at UCLA; Jennifer S. Polson, Heart BD2K Center at UCLA; Shashank Khanna, Heart BD2K Center at UCLA; Austin Nasso, Heart BD2K Center at UCLA; Quan Cao, Heart BD2K Center at UCLA; Chunyu Guo, Heart BD2K Center at UCLA; David Liem, Heart BD2K Center at UCLA; Jie Wang, Heart BD2K Center at UCLA; Anders O. Garlid, Heart BD2K Center at UCLA; Vincent Kyi, Heart BD2K Center at UCLA; Peipei Ping, Heart BD2K Center at UCLA*

A wealth of biomedical information is contained within published literature and databases; however, much of these data are scattered and fragmented, requiring a significant investment of time to gather and comprehend. Gene Wiki is a portal within Wikipedia that enables the aggregation and translation of gene data to knowledge through collective curation. We aimed to develop a framework to harness this crowdsourcing potential toward generating and improving public access to cardiovascular knowledge. We established standardized criteria to score 556 human heart mitochondrial, 64 heart contractile, and 26 coronary artery disease (CAD) risk gene/protein pages based on three core attributes: (i) gene/protein structure, (ii) biological function, and (iii) clinical significance. To date, 422 of the 646 total selected proteins (65%) have been curated with a complete entry; this represents 21 of the most investigated molecules in cardiovascular medicine and the entire set of CAD genetic risk factors. To expedite the transformation of data into knowledge, we have been developing an online tool that integrates the research and writing aspects of biocuration in one user interface, enhancing users' capability to extract information from literature and to gain new knowledge. Our tool, BioComposer, is hosted on the Amazon Elastic Compute Cloud (EC2) web service and constructed using the programming language Node.js and database program MongoDB. The source code is available on GitHub (https://github.com/UCLA-BD2K/GeneWiki). Version 1.0 is currently online (genewiki-env.us-west-2.elasticbeanstalk.com) and highlights two primary features: (i) a search engine that enables users to find relevant text data (e.g., scientific articles) and manage citations; and (ii) a rich text editor that enables users to create and edit text data (e.g., Gene Wiki pages). Overall, our efforts promote the identification and the comprehension of existing cardiovascular information, driving scientific discovery and building new knowledge.

Pathfinder: Visual Analysis of Paths in Biological Networks

***Alexander Lex, University of Utah**; Christian Partl, Graz University of Technology, Austria; Samuel Gratzl, Johannes Kepler University Linz, Austria; Marc Streit, Johannes Kepler University Linz, Austria; Anne Mai Wassermann, Pfizer; Hanspeter Pfister, Harvard University; Dieter Schmalstieg, Graz University of Technology, Austria*

The analysis of paths in graphs is highly relevant in biology, with applications in pathway analysis, protein interaction networks, connectomics etc. Typically, path-related tasks are performed in node-link layouts. Unfortunately, graph layouts often do not scale to the size of many real world networks. Also, biological networks commonly are multivariate, i.e., contain rich attribute sets associated with the nodes and edges. In the pathway example, these attributes could be omics data associated with the genes (nodes), capturing, for example, gene expression, copy number variation, mutation information etc. These attributes are often critical in judging paths, but directly visualizing attributes in a graph layout exacerbates the scalability problem. We presented visual analysis solutions dedicated to path-related tasks in large and highly multivariate graphs. We show that by focusing on paths, we can address the scalability problem of multivariate graph visualization, equipping analysts with a powerful tool to explore large graphs. We introduce Pathfinder, a technique that provides visual methods to query paths, while considering various constraints. The resulting set of paths is visualized in both a ranked list and as a node-link diagram. For the paths in the list, we display rich attribute data associated with nodes and edges, and the node-link diagram provides topological context. The paths can be ranked based on topological properties, such as path length, and scores derived from attribute data. Pathfinder is designed to scale to graphs with tens of thousands of nodes and edges by employing strategies such as incremental query results. We demonstrate Pathfinder's fitness for use in a signaling pathway network in the context of gene expression and copy number data.

Pathfinder is web based and open source. A demo and Details about the technique can be found at: http://www.caleydo.org/publications/2016_eurovis_pathfinder/

GEN3VA: Aggregation and Analysis of Gene Expression Signatures From Related Studies

Avi Ma'ayan, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; *Gregory Gundersen, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Nicolas Fernandez, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Caroline D. Montiero, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Kathleen M. Jagodnik, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai; Anders B. Dohlman, BD2K-LINCS DCIC, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai*

Genome-wide gene expression profiling of mammalian cells is becoming a staple of many published biomedical and biological research studies. Such data is deposited into data repositories such as the Gene Expression Omnibus (GEO) for potential reuse. However, these databases currently do not provide simple strategies to systematically analyze collections of related studies. Here we present GENE Expression and Enrichment Vector Analyzer (GEN3VA), a web-based system that enables the integrative analysis of aggregated collections of tagged gene expression signatures identified and extracted from GEO. Each tagged collection of signatures is presented in a report that consists of heatmaps of the differentially expressed genes; principal component analysis of all signatures; enrichment analysis with several gene set libraries across all signatures, which we term enrichment vector analysis; and global mapping of small molecules that are predicted to reverse or mimic each signature in the aggregate. We demonstrate how GEN3VA can be used to identify common molecular mechanisms of aging by analyzing tagged signatures from 244 studies that compared young vs. old tissues in mammalian systems. In a second case study, we collected 86 signatures from treatment of human cells with dexamethasone, a glucocorticoid receptor (GR) agonist. Our analysis confirms consensus GR target genes and predicts potential drug mimickers. GEN3VA can be used to identify, aggregate, and analyze themed collections of gene expression signatures from diverse but related studies. Such integrative analyses can be used to address concerns about data reproducibility, confirm results across labs, and discover new collective knowledge by data reuse. GEN3VA is an open-source web-based system that is freely available at: http://amp.pharm.mssm.edu/gen3va.

A MACE2K Text Mining Tool to Extract the Impact of Genomic Anomalies on Drug Response

A.S.M. Ashique Mahmood, Department of Computer and Information Science, University of Delaware; *Shruti Rao, Innovation Center for Biomedical Informatics, Georgetown University; Peter McGarvey, Innovation Center for Biomedical Informatics, Georgetown University, Protein Information Resource, Georgetown University Medical Center; Cathy Wu, Department of Computer and Information Science, University of Delaware, Center for Bioinformatics and Computational Biology, University of Delaware, Protein Information Resource, Georgetown University Medical Center; Subha Madhavan, Innovation Center for Biomedical Informatics, Georgetown University, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center; K. Vijay-Shanker, Department of Computer and Information Science, University of Delaware*

Rapidly evolving genomics tools and technologies have contributed to a dramatic rise in the volume and complexity of cancer precision medicine literature. This makes it challenging for an oncologist to search, perceive and use the information for personalized treatment of patients. Moreover, manual curation from literature is time-consuming and expensive. Hence, there is an urgent need to automatically extract information from literature to assist curators and clinical researchers. In this work, we have developed a text mining system to address this need.

Our natural language processing (NLP) tool detects different entities from PubMed abstracts and extracts the entity relationships that indicate the impact of genomic anomalies on cancer therapeutics. Our tool accounts for a variety of ways such relationships can be described in text. To assist with such extraction of information, we have extended and repurposed multiple in-house and public text mining systems including DiMeX, miRiaD and PubTator. Moreover, we extract additional details about the study to help experts determine the quality of evidence of the study. The NLP tool produces output in both JSON and BioC formats to facilitate data exchange and integration of text mining results. Additionally, the BioC representation can enable the use of existing curation tools for expert validation and ranking. All extracted results are stored in a database and are available for curators and clinical researchers via an interactive web interface. The extracted information relating genomic anomalies to drug responses will enable researchers to readily generate hypotheses for new precision medicine based clinical trials.

Faster and Better Metadata Authoring Using CEDAR's Value Recommendations

Marcos Martinez-Romero, Stanford University; *Martin J. O' Connor, Stanford University; Maryam Panahiazar, Stanford University; Debra Willrett, Stanford University; Attila L. Egyedi, Stanford University; John Graybeal, Stanford University; Mark A. Musen, Stanford University*

In biomedicine, good metadata is crucial to finding experimental datasets, to understand how experiments were performed, and to reuse data to conduct new analyses. Despite the growing number of efforts to define guidelines and standards to describe biomedical experiments, the impediments to creating accurate, complete, and consistent metadata are still considerable. Authoring good metadata is a tedious and time-consuming task that biomedical scientists tend to avoid.

The Center for Expanded Data Annotation and Retrieval (CEDAR) is developing novel methods and tools to simplify the process by which investigators annotate their experimental data with metadata. The CEDAR Workbench (cedar.metadatacenter.net) is a set of Web-based tools for the acquisition, storage, search, and reuse of metadata templates. As a step towards decreasing authoring time while increasing metadata quality, we have enhanced the CEDAR Workbench with value recommendation capabilities.

Our system identifies common patterns in the CEDAR metadata repository, and generates real-time suggestions for filling out metadata acquisition forms. These suggestions are context-sensitive, meaning that the values predicted for a particular field are generated and ranked based on previously entered values. Our value recommendation approach supports both free-text values and terms from ontologies and controlled terminologies. We discuss some of the challenges that have arisen while implementing our approach, and our strategies for making this capability useful to the end users of CEDAR. We demonstrate CEDAR's intelligent authoring capabilities using metadata from the Gene Expression Omnibus (GEO), and show how the technology that we are developing leverages existing metadata to make the authoring of high-quality metadata a manageable task.

New Algorithms for RNA-seq and ChIP-seq Data Compression

Olgica Milenkovic, University of Illinois at Urbana-Champaign; *Vida Ravanmihr, University of Illinois; Zhiying Wang, University of California; Minji Kim, University of Illinois*

We report on smallWig and ChiPWig, two new software solutions for RNA-Seq and ChiP-Seq data compression. Our new low-rate compression methods are especially designed for RNA- and ChIP-seq data, which contains real valued entries with patterns that cannot be efficiently compressed by general purpose compression software.

Our approaches are based on statistical modeling of position and expression values, and source coding techniques, which include transform coding, differential coding and arithmetic coding for integers and correlated real numbers. An additional compression technique, known as context-tree weighting, was used to achieve ultrahigh compression rates needed for archival storage applications

We tested our compression methods on different RNA-Seq and ChIP-seq data generated by the ENCODE project. The results reveal that the new methods offer, on average, a 10/20-fold decrease in file size compared to bigWig, while providing the same summary statistics and random access features. Compression and decompression times are comparable to those of wig formats.

kBOOM! Intelligent Merging of Different Disease Terminologies

Chris Mungall, Lawrence Berkeley National Laboratory; *Sebastian Koehler, Charite; Peter Robinson, The Jackson Laboratory; Nicole Vasilesky, The Ohio State University; Ian Holmes, University of California, Berkeley; Melissa Haendel, The Ohio State University*

Ontologies allow us to create a map of our biological universe, providing shared landmarks on which to attach data, enabling interoperability and enhancing findability. However, much of the territory is still unknown and our maps are fragmented, often covering different scales or perspectives. This is especially true in the part of our universe concerning diseases, where we must stitch together different artifacts to achieve a unified picture of the terrain. For example, the Online Mendelian Inheritance in Man (OMIM) resource provides a comprehensive list of Mendelian diseases; the NCI Thesaurus is authoritative for neoplasm classification; Orphanet classifies rare diseases; and the Disease Ontology classifies rare, common, and infectious diseases, but gaps remain. Many resources such as Bioportal provide mappings between landmarks on these different maps, but these are frequently vague or inconsistent and thus there is no resource that performs the weaving into a single cohesive, consistent unified map.

We have developed a method that takes multiple resources as input, together with mappings, and generates a single coherent ontology, with equivalent diseases from different resources merged into a single entry, and logically consistent relationships connecting diseases together, providing a merged resource. The merged disease ontology can be used for use in metadata descriptions for a wide variety of data, ensuring interoperability. The method is called kBOOM (Bayesian OWL Ontology Merging) and uses a combination of logical reasoning and probabilistic methods. The method is applicable to a variety of terminologies and ontologies and will form part of the suite of web based tools available as part of the new BD2K INtelligent Concept Assistant (INCA) project, https://github.com/INCATools/intelligent-concept-assistant that will provide a means for scientists to collect together and assemble terms and standards for use in describing data.

Using Crowds to Design Biological Network Visualizations

T.M. Murali, Department of Computer Science, Virginia Tech; *Divit P. Singh, Department of Computer Science, Virginia Tech; Kurt Luther, Department of Computer Science, Virginia Tech*

Biologists often perform experiments whose results generate large quantities of data, such as interactions between molecules in a cell, that are best represented as networks (graphs). To visualize these networks and communicate them in publications, biologists must manually position the nodes and edges of each network to reflect their real-world physical structure. This process does not scale well, and graph layout algorithms lack the biological underpinnings to offer a viable alternative. We present GraphCrowd, a crowdsourcing system that leverages human intelligence and creativity to design layouts of biological network visualizations. GraphCrowd provides design guidelines, abstractions, and editing tools to help novice workers perform like experts.

We evaluated GraphCrowd with two experiments. The first found that crowdsourced layouts of real biological networks were as good as or better than layouts designed by expert biologists, and significantly better than a popular graph drawing algorithm. A second experiment found that crowds provided quality ratings of network layouts that were similar to an expert biologist, suggesting that GraphCrowd can both create and identify high-quality layouts.

Our contributions include: 1) novel techniques for generating and evaluating scalable, high-quality biological network visualizations via novice crowdsourcing, 2) experiments providing empirical evidence of the benefits of these techniques compared to expert and algorithmic baselines, 3) implications for crowdsourcing complex design work, including other types of network visualizations, and 4) the GraphCrowd system itself, which we will release as open-source software.

ADAM Enables Distributed Analyses Across Large-Scale Genomic Datasets

Frank Nothaft, University of California, Berkeley; *Arun Ahuja, Icahn School of Medicine at Mount Sinai; Timothy Danford, AMPLab, University of California, Tamr; Michael Heuer, AMPLab, University of California; Jey Kottalam, AMPLab, University of California; Matt Massie, AMPLab, University of California; Audrey Musselman-Brown, Genome Informatics Lab, University of California; Beau Norgeot, Genome Informatics Lab, University of California; Ravi Pandya, Microsoft Research; Justin Paschall, AMPLab, University of California; Jacob Pfeil, Genome Informatics Lab, University of California; Hannes Schmidt, Genome Informatics Lab, University of California; Eric Tu, AMPLab, University of California; John Vivian, Genome Informatics Lab, University of California; Ryan Williams, Icahn School of Medicine at Mount Sinai; Carl Yeksigian, GenomeBridge; Michael Linderman, Icahn School of Medicine at Mount Sinai; Jeff Hammerbacher, Icahn School of Medicine at Mount Sinai; Uri Laserson, Icahn School of Medicine at Mount Sinai, Cloudera, Inc.; Gaddy Getz, Broad Institute; David Haussler, Genome Informatics Lab, University of California; Benedict Paten, Genome Informatics Lab, University of California; Anthony D. Joseph, AMPLab, University of California; David A. Patterson, AMPLab, University of California; ASPIRE Lab, University of California*

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 abstract presents recent updates to ADAM, an Apache 2 licensed library built on top of the popular Apache Spark distributed computing framework. We are using ADAM and the Toil workflow management system (Apache 2 licensed) to recall the Simons Genome Diversity project dataset against the GRCh38 build of the human reference genome. Because ADAM is designed to allow genomic analyses to be seamlessly distributed across large clusters, we achieve a 3.5× improvement in end-to-end variant calling latency and a 66% cost improvement over current toolkits, without sacrificing accuracy.

GTRAC: Fast Retrieval From Compressed Collections of Genomic Variants

Idoia Ochoa, Stanford University

The dramatic decrease in the cost of sequencing has resulted in the generation of huge amounts of genomic data, as evidenced by projects such as the UK10K and the Million Veteran Project (MVP), with the number of sequenced genomes ranging in the order of 10K to 1M. Due to the large redundancies among genomic sequences of individuals from the same species, most of the medical research deals with the variants in the sequences as compared with a reference sequence, rather than with the complete genomic sequences. Consequently, millions of genomes represented as variants are stored in databases. These databases are constantly updated and queried to extract information such as the common variants among individuals or groups of individuals. Previous algorithms for compression of this type of databases lack efficient random access capabilities, rendering querying the database for particular variants and/or individuals extremely inefficient, to the point where compression is often relinquished altogether.

We present a new algorithm for this task, called GTRAC, which achieves significant compression ratios while allowing fast random access over the compressed database. For example, GTRAC is able to compress a H. Sapiens dataset containing 1092 samples in 1.1GB (compression ratio of 160), while allowing for decompression of specific samples in less than a second and decompression of specific variants in 17ms. GTRAC uses and adapts techniques from information theory, such as a specialized Lempel-Ziv compressor, and tailored succinct data structures.

A Standards-Based Model for Metadata Exchange

Martin O'Connor, Stanford University; *Marcos Martínez-Romero, Stanford University; Attila L. Egyedi, Stanford University; Debra Willrett, Stanford University; John Graybeal, Stanford University; Mark A. Musen, Stanford University*

There has been a dramatic increase in the availability of biomedical data sets derived from scientific experiments. High-quality descriptive metadata is seen as essential to facilitate the discovery and interpretation of these data sets. The biomedical community has developed templates to describe metadata for a variety of experiment types, providing a strong foundation for the development of a large number of public metadata repositories.

Unfortunately, these templates rarely share common structure or semantics. Moreover, biomedical repositories usually require proprietary submission formats that are often loosely connected to underlying template specifications. Crucially, these formats typically lack standard mechanisms for semantically annotating the metadata in templates. These difficulties combine to ensure that most metadata submissions have weak semantic content.

A key shortcoming is the absence of an interoperable format for metadata exchange. Driven by the goals of the Center for Expanded Data Annotation and Retrieval (CEDAR) (metadatacenter.org), we have developed such a format. We created a lightweight standards-based template model that provides both a structural specification of metadata and principled interoperation with controlled terminologies and Linked Open Data.

In addition to semantically marking up templates themselves, the model supports ontology-based value constraints to ensure that metadata conforming to these templates are linked to controlled terminologies. The model also provides mechanisms to support template composition, with the aim of increasing reuse of metadata fragments across templates.

We developed an implementation of this model using the Web-centric JSON format. The associated JSON Schema (json-schema.org) and JSON-LD (json-ld.org) specifications provide standard technologies to represent the structural aspects of CEDAR's template model and the linkage to semantic technologies.

We created a Web-based ecosystem driven by the model to provide an end-to-end workflow for metadata acquisition and management. We released a public alpha version of the system in September 2016 (cedar.metadatacenter.net).

Mining Electronic Health Records for Possible Drug Repositioning Opportunities

David Page, University of Wisconsin-Madison; *Charles Kuang, University of Wisconsin-Madison; James Thomson, Morgridge Institute; Michael Caldwell, Marshfield Clinic; Peggy Peissig, Marshfield Clinic; Ron Stewart, Morgridge Institute*

Computational Drug Repositioning (CDR) is the task of discovering potential new uses for existing drugs by mining large-scale heterogeneous drug-related data sources. Leveraging the patient-level temporal ordering information between numeric physiological measurements and various drug prescriptions provided in Electronic Health Records (EHRs), we propose a Continuous Self-controlled Case Series (CSCCS) model for CDR. As an initial evaluation, we look for drugs that can control Fasting Blood Glucose (FBG) level in our experiments. Applying CSCCS to the Marshfield Clinic EHR, well-known drugs that are indicated for controlling blood glucose level are rediscovered. Furthermore, some drugs with recent literature support for potential blood glucose level control are also identified.

Dynamic Control Models for Strategic Interaction

John Pearson, Duke University; *Shariq N. Iqbal, Duke University; Caroline B. Drucker, Duke University; Michael L. Platt, University of Pennsylvania*

The ecological niches occupied by most organisms, including humans, are both dynamic and uncertain, requiring that actions be taken in real time and modified in response to changing circumstances. However, most studies of dynamic decision-making to date have explored either repeated trials of the same task under slowly changing circumstances (e.g., bandit problems) or interactions between agents that take place in a restricted action space (e.g., prisoner's dilemma). Here, we examine data from repeated trials of a real-time strategic interaction with continuous freedom of movement. We trained monkeys to play a competitive task in which the goal of one (the "shooter") was to move a colored dot (the "ball") from the left to right side of a computer monitor using joystick input. The goal of the second monkey (the "goalie") was to block the dot by moving a vertical line along the right-hand side of the screen to intercept it. Thus, each player controlled an avatar with at least one continuous degree of freedom, in principle allowing for dynamic coupling between the two in real time.

We analyzed these data using time series methods borrowed from the machine learning and optimal control literatures. We modeled each player's avatar as a state space model with control input dependent on the dynamics of both players' behavior. That is, the shooter's control action was modeled as a filtered sum of his own and the goalie's past trajectories (and vice-versa for the goalie). The model produces a set of filters and time series that can be used for subsequent analysis of neural data in terms of dynamic control signals derived from behavior.

Automatic Discovery and Processing of EEG Cohorts From Clinical Records

Joseph Picone, Temple University; *Iyad Obeid, Temple University; Sanda Harabagiu, University of Texas at Dallas*

Decision support systems in healthcare can leverage vast archives of electronic medical records if high performance automated data wrangling can be achieved. EMRs can include unstructured text, temporally constrained measurements (e.g., vital signs), multichannel signal data (e.g., EEGs), and image data (e.g., MRIs). Our focus is the automatic interpretation of a clinical EEG Big Data resource known as the TUH EEG Corpus (TUH EEG). There are four major aims in this project: (1) automatically recognize and time-align events in EEG signals, (2) automatically recognize critical concepts in the EEG reports, (3) automatic patient cohort retrieval, and (4) evaluation and analysis of the results of the patient cohort retrieval.

We have developed two demonstrations of our cohort retrieval technology: a Multi-Modal EEG Patient Cohort Retrieval system called MERCuRY an acronym for Multi-modal EncephalogRam patient Cohort discoverY) and an EEG signal visualization tool. We have developed novel methods of identifying in the EEG reports the EEG activities, EEG events and patterns as well as their attributes. In addition to the EEG-specific medical concepts, we have also identified through our methods all medical concepts that describe the clinical picture and therapy of the patients. We have validated the usefulness of the patient cohort identification system by collecting feedback from clinicians and medical students.

Identification of the type and temporal location of EEG signal events such as spikes or generalized periodic epileptiform discharges in the EEG signal are critical to the interpretation of an EEG. We have developed a high performance event detection system that integrates hidden Markov models for event detection and deep learning for postprocessing. Our performance on a standard clinical event recognition task has improved to 91.4% sensitivity with an 8.5% specificity, which is approaching the level of performance required by clinicians.

EEG Event Detection Using Deep Learning

Joseph Picone, Temple University; *Meysam Golmohammadi, Temple University; Saeedeh Ziyabari, Temple University; Silvia Lopez, Temple University; Elliott Krome, Temple University; Matthew Thiess, Temple University; Scott Yang, Temple University; Iyad Obeid, Temple University*

Automated data wrangling for physiological signals commonly found in healthcare, such as electroencephalography (EEG) signals, requires identification and localization of events in time and/or space. Deep learning systems, which can achieve impressive levels of performance on such sequential data, require vast amounts of annotated data, often referred to as Big Data, to achieve high performance. Normalization of data with respect to annotation standards, recording environments, equipment manufacturers and even standards for clinical practice, must be accomplished for technology to be clinically relevant. Many decision support systems in healthcare could be successfully automated if such data resources existed. In this demonstration, we will introduce a high performance system for EEG event detection that enables keyword searches of a large archive of EEG signal data and can be used for automated data wrangling.

This baseline system integrates low-level event processing using hidden Markov models with higher-level event interpretation using deep learning. Self-training has been used to facilitate semi-automated annotation of training data. A supervector approach to feature extraction coupled with Principle Component Analysis has been used for spatial localization of key events such as seizures. A stacked denoising autoencoder has been used to postprocess event hypotheses and convert these into term hypotheses. A new term-based scoring process, popular in evaluation of audio-based keyword search systems, is introduced and compared to several existing evaluation methodologies. Several new research resources based on the TUH EEG Corpus, developed to enable this research, will also be introduced. A Python-based demonstration of an EEG visualization tool will be provided as well.

Scalable EEG Interpretation Using Deep Learning and Schema Descriptors

Joseph Picone, Temple University; *Iyad Obeid, Temple University; Sanda Harabagiu, The University of Texas at Dallas*

This project addresses a critical market gap in EEG technology – real-time seizure detection for intensive care unit (ICU) and epilepsy monitoring unit (EMU) applications. The ability to auto-scan EEGs and predict seizures in advance will be a transformational clinical technology. Existing products that seek to increase accuracy and productivity via automatic analysis are limited by high rates of false positives, overwhelming healthcare providers with misleading information. Recently, deep learning systems have made tremendous progress in delivering powerful solutions from loosely transcribed data, due to rapid advances in low-cost, highly-parallel computational infrastructure and powerful machine learning algorithms. However, these techniques, which have been very successful in fields such as speech recognition and image understanding because large amounts of training data are available, have yet to be applied to biomedical signals such as EEGs due to a lack of big data resources. Hence, a major goal of this supplement is to complete a pilot study demonstrating the feasibility of applying these big data techniques to biomedical signals.

There are four specific aims: (1) automatic labeling of the TUH EEG Corpus for seizure events; (2) application of deep learning sequential modeling techniques for EEGs; (3) defining Hierarchical epileptiform Activity Descriptors (HAD) for EEGs; and (4) automated Tagging of HADs in medical texts.

The proposed work will produce several novel resources, including (1) a labeled subset of the publicly available TUH EEG Corpus that can support seizure detection research; (2) software tools that apply state of the art sequential modeling systems based on deep learning to biomedical signal processing; (3) The Hierarchical epileptiform Activity Descriptor (HAD) schema, which will provide a scalable solution for representing and annotating epileptiform activities, including those related to seizure prediction, both in biomedical and clinical documents; (4) automatic methods for generating HAD schema tags on EEG reports.

Integrative LINCS (iLincs): Connecting Diseases, Drugs, and Mechanisms of Actions

Marcin Pilarczyk, University of Cincinnati; *Mehdi Fazel Najafabadi, University of Cincinnati; Michal Kouril, University of Cincinnati; Naim Mahi, University of Cincinnati; Nicholas Clark, University of Cincinnati; Shana White, University of Cincinnati; Mark Bennett, University of Cincinnati; Wen Niu, University of Cincinnati; John Reichard, University of Cincinnati; Juozas Vasiliauskas, University of Cincinnati; Jarek Meller, University of Cincinnati; Mario Medvedovic, University of Cincinnati*

iLINCS (Integrative LINCS) is an integrative web platform for analysis of omics data and signatures of cellular perturbations. The portal consists of biologists-friendly user interfaces for finding and analyzing datasets and signatures, backend databases with a large collection of datasets (>3,000), pre-computed signatures (>200,000) and their connections (>2109). The portal integrates R analytical engine via several R tools for web-computing (rserve, opencpu, shiny, rgl) and other public domain web tools and open-source applications (eg FTreeView, Enrichr, L1000CDS2) into a coherent web platform for omics data analysis. Analytical tools are organized into three interconnected analytical workflows.

The “Dataset” workflow facilitates comprehensive analysis of primary omics datasets. In a typical use case, the user starts with an omics dataset of interest (eg GEO dataset corresponding to a disease of interest), performs differential gene expression analysis to construct the signature of the disease. Performs enrichment, pathway and network analysis of differentially expressed genes. Identifies “connected” drug signatures that can implicate a potential therapeutic agent for the disease.

The “Signatures” workflow facilitates “connectivity analysis” with a large collection of pre-computed signatures that include LINCS drug perturbation signatures, ENCODE transcription factor binding signatures and a library of “disease related signatures” extracted from public domain omics datasets. User can either select one or more pre-computed signatures, or upload their own signatures to use in the analysis. One of the use-cases involves uploading a custom disease signature, identifying the connected LINCS chemical perturbagen signatures which can then provide putative agents for treating the disease.

The “Genes” workflow starts with a user supplied list of genes which are then used to query and analyze primary data and pre-computed signatures.

The portal can be accessed freely and does not require user registration (http://ilincs.org).

NetLINCS: Correlation of Chemical Perturbagen and Signatures to Identify Biological Targets

John Reichard, University of Cincinnati; *J.F. Reichard; M. Bennett; M. Pilarczyk; W. Niu; M. Medvedovic*

Data science has the potential to transform fields of research, such as toxicological. For example, rather than looking at biological responses to individual chemicals, data science approaches can be used to identify biological response patterns that are shared within or between groups of chemicals or biological models. High-throughput bioassays, such as those generated by the Library of Network-Based Cellular Signatures (LINCS) consortium (http://www.lincsproject.org/) provide an opportunity to identify such patterns. Here, we have used the NetLINCs perturbation signature-based analytical framework (PMID: 24039560) to identify key pathway differences across several cancer cell lines. The approach leverages LINCS L1000 gene perturbation signatures; however, the methodology is also applicable to many other signature types (e.g. proteomic, RNA-Seq). The methodology is based on correlating all chemical perturbation signatures with gene knockdown signatures. Receiver operating characteristic (ROC) curves are then generated that use known chemical-target relationships as a binary discriminator to distinguish true positive and false positive correlations. The area under the curve (AUC) is calculated for chemical-target pairs to identify highly predictive (high true positive, low false negative rates) correlations. Difference in AUC among cell lines is then used to identify target-related pathway differences that distinguish cell models. From these results, the biological basis for differences among cell types can be explored. On approach is using the iLINC web portal (http://www.ilincs.org/ilincs/#/) to characterize signature correlation patterns, perform gene enrichment analysis and develop pathway-based understandings. The methodology is described, and examples are provided to illustrate how the NetLINCS approach can be used to obtain insight into chemical mode of action. In theory, these mechanistic relationships could be extended to identify chemicals with highly concordant signatures for which mechanistic targets are characterized or not available in the public domain. Funding: NIH Common Fund as part of the LINCS project (U54 HL127624).

Compressive Structural Bioinformatics: Large-Scale Analysis and Visualization of the PDB Archive

Peter Rose, University of California, San Diego; *Anthony R. Bradley, University of California, San Diego; Alexander S. Rose, University of California, San Diego; Yana Valasatava, University of California, San Diego; Jose M. Duarte, University of California, San Diego; Andreas Prlić, University of California, San Diego*

We are developing compressed 3D molecular data representations and workflows ("Compressive Structural Bioinformatics") to speed up visualization and mining of 3D structural data by one or more orders of magnitude. Our data representations allow scanning and analyzing the entire PDB archive in minutes or visualizing structures with millions of atoms.

Compact data representation - Existing text-based file formats for macromolecular data (PDBx/mmCIF) are slow to parse. To address this bottleneck, we have developed the Macromolecular Transmission Format (MMTF) (http://mmtf.rcsb.org/) that offers 75% compression over the standard mmCIF format and is over an order of magnitude faster to parse. MMTF files are easy to decode using our open source libraries, and hence the format has been rapidly adopted by the community, including the following 3D viewers: 3Dmol.js, iCn3D (NCBI), ICM, Jmol, PyMol, and bioinformatics libraries: BioJava, and Biopython.

High-performance web-based visualization - The small individual file size and high parsing speed enables high performance web-based visualizations. We have seen a greater than 20x speedup over mmCIF in loading of PDB entries from sites across the USA, Europe, and Asia. Using MMTF and NGL, a highly memory-efficient WebGL-based viewer, even the largest structures in PDB can be visualized on a smartphone.

High-performance distributed parallel workflows - MMTF enables scalable Big Data analysis of 3D macromolecular structures using distributed, in-memory, parallel processing frameworks such as Apache Spark. For example, we can extract all ligands from the PDB using MMTF with Apache Spark in about 3 minutes. In contrast, using the mmCIF format, the same task takes several hours.

The MMTF file format enables a paradigm change for structural bioinformatics applications. It is now possible to store the entire PDB in memory to eliminate I/O bottlenecks, to rapidly visualize large structures over the web, and to perform distributed parallel processing.

Scientific Reproducibility Using the Provenance for Clinical and Healthcare Research Framework

Satya Sahoo, Case Western Reserve University; *Joshua Valdez, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University; Michael Rueschman, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University; Matthew Kim, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University; Susan Redline, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University*

Scientific reproducibility is key to scientific progress as it allows the research community to build on validated results, protect patients from potentially harmful trial drugs derived from incorrect results, and reduce wastage of valuable resources. The National Institutes of Health (NIH) recently published a systematic guideline titled "Rigor and Reproducibility" for supporting reproducible research studies, which has also been accepted by several scientific journals. These journals will require published articles to conform to these new guidelines. Provenance metadata describes the history or origin of data and it has been long used in computer science to capture metadata information for ensuring data quality and supporting scientific reproducibility. In this paper, we describe the development of Provenance for Clinical and Healthcare Research (ProvCaRe) framework together with a provenance ontology to support scientific reproducibility by formally modeling a core set of data elements representing details of research study. We extend the PROV Ontology (PROV-O), which has been recommended as the provenance representation model by World Wide Web Consortium (W3C), to represent both: (a) data provenance, and (b) process provenance. The ProvCaRe ontology is being used to extract provenance metadata to generate Resource Description Framework (RDF) provenance graphs from biomedical text to characterize provenance metadata described in published scientific studies that can be used for replicating scientific studies. We describe the results of analyzing 50 research publications using the ProvCaRe framework.

Computational Tools and Resources for LINCS Proteomics Data

Behrouz Shamsaei, University of Cincinnati; *Szymon Chojnacki; Vagisha Sharma; Jeremy Muhlich; Dusica Vidovic; Wen Niu; Jake Jaffe; Mario Medvedovic; Jarek Meller*

Versatile, re-usable interfaces are required to facilitate integration of LINCS proteomics data into the tools and platforms being developed by the LINCS/BD2K community. Protein Line Notation or PLN Converter (http://eh3.uc.edu/pln) is developed to enable consistent mapping and annotation of peptide probes, including protein and pathway mapping, and harmonize proteomic data sets across different centers and assays. piLINCS (pilincs.org) has been designed to provide both interactive and API level access to reduced phosphoproteomic (P100) and global chromatin (GCP) pertubagen profiles generated by the Broad Institute and made available through Panorama, and as a basis for further integration of LINCS proteomic data.

SAP – A CEDAR-Based Pipeline for Semantic Annotation of Biomedical Metadata

Ravi Shankar, Stanford University; *Marcos Martinez-Romero, Stanford University; Martin J. O'Connor, Stanford University; John Graybeal, Stanford University; Purvesh Khatri, Stanford University; Mark A. Musen, Stanford University*

The exponential growth in the volume of biomedical data held in public data repositories has created tremendous opportunity to evaluate novel research hypotheses in silico. But such search and analysis of disparate data presupposes a consistent semantic representation of the metadata that annotate the research data. Semantic grouping of data is the cornerstone of efficient searches and meta-analyses. Existing metadata are either granularly defined as tag–value pairs (e.g., sample organism=""homo sapiens"") or implicitly found in long textual descriptions (e.g., in a study design overview). Current practice is to manually map metadata strings to ontological terms before any data analysis can begin. But manual semantic annotation is time-consuming and requires domain and ontology expertise, and therefore may not scale with metadata growth.

Under the umbrella of the Center for Enhanced Data Annotation and Retrieval (CEDAR) metadata enrichment effort, we are building the Semantic Annotation Pipeline (SAP), which automates semantic annotation of biomedical data stored in public data repositories. The pipeline has two major segments: 1) Reformat the metadata stored in the data repository to CEDAR JSON-LD format, using templates created in the CEDAR repository, and 2) Add semantic annotations to the CEDAR formatted metadata. We are employing Apache's UIMA ConceptMapper to efficiently map metadata text segments to ontology terms.

We are using the GEO microarray data repository to build and evaluate SAP. Our initial focus is on annotating a specific set of experiment metadata including experiment design, and sample characteristics such as organism, disease, and treatment. We plan to evaluate the SAP annotations against manually curated data, including GEO datasets found in the GEO repository. We intend to show that SAP can ease the process of semantic annotation of metadata, and the enriched metadata can support efficient search and meta-analyses of biological and biomedical data.

Using Eye Tracking to Enhance Usability of Big Data in Cancer Precision Medicine

Vishakha Sharma, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; *Robert A. Beckman, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; Shruti Rao, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; Allan Fong, National Center for Human Factors in Healthcare, MedStar Health; A. Zachary Hettinger, National Center for Human Factors in Healthcare, MedStar Health; Simina M. Boca, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; Peter B. McGarvey, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; Raj M. Ratwani, National Center for Human Factors in Healthcare, MedStar Health; Subha Madhavan, Innovation Center for Biomedical Informatics, Georgetown University Medical Center*

The success of healthcare and clinical research software applications, especially those built for clinicians and scientists, is dependent on understanding the complex cognitive processes of the intended users and their workflow processes. The potential power of precision medicine can be impeded by poor usability of information systems. We leverage human factors engineering to make the complex molecular and clinical information extracted from the literature more accessible and actionable to clinicians.

As part of our software suite MACE2K (Molecular and Clinical Extraction to Knowledge) we have developed a Natural Language Processing (NLP) tool to automatically extract entities and relations from precision medicine literature. Once extracted, this information must be synthesized and summarized for use by clinicians who generally have little time to inspect the evidence. We have built prototype user interfaces for MACE2K to maximize comprehension of evidence from clinical actionable biomarkers including a concise recommendation for action within the context of a level of evidence framework.

To better understand how the clinician visually processes information on the interfaces and how the interfaces support the cognitive processes of the clinician (e.g. reasoning and decision-making), we use a unique methodology consisting of an eye-tracking device with talk-aloud verbal protocol. We defined different areas of the user interfaces as areas of interest (AOIs). The AOIs serve as meaningful aspects of the interface to study the eye movement data patterns in conjunction with the verbal interpretations provided by the clinicians. This methodology allows the clinicians and clinical researchers to more clearly identify which aspects of the interface support reasoning and decision-making and which aspects of the interface are less optimal.

The application of human factors engineering to our evidence based information on clinically actionable biomarkers has tremendous potential and will make this information more easily usable in real clinical applications.

Big Data Contrast Mining for Genetic Distinctions Between Disease Subtypes

Matt Spencer, University of Missouri; *Chi-Ren Shyu*

Diseases with complex etiologies require complex methods to decipher them. GWAS is effective, but has two major limitations: it inherently fails to account for relationships between SNPs, and most procedures treat affected cases as a homogeneous group. However, complex diseases are often conglomerations of multiple diseases that involve interacting SNPs. Thus, there is a need for novel methods that examine SNP combinations and contrast disease subgroups.

Assessing SNP combinations is a Big Data problem; genotype datasets comprise millions of SNPs. To address this, we utilize data mining tools implemented in Apache Spark, the in-memory Big Data computing framework. Combinations of SNPs are generated using Frequent Pattern Mining (FPM); this data-mining algorithm considers SNP prevalence in the affected population while isolating combinations to examine.

Additionally, it is essential to identify genetic differences between disease-related subgroups, allowing treatments to be developed addressing the specific nature of the subtypes. We accomplish this using Contrast Mining: affected people are divided into clinically relevant groups, using phenotypes or biomarkers, then combination prevalence is calculated separately for each group. Contrast Mining then systematically compares groups, identifying SNP combinations heavily favoring one group.

Autism is a complex disorder known for its extreme diversity, suggesting a multitude of etiologies. We used these Big Data tools to contrast two phenotypically distinct autism subtypes for which the underlying genetic mechanisms distinguishing them are unknown. We identified 8 individual SNPs and 23 SNP combinations differentiating the subtypes. We plan to apply these in-memory computing and exploratory analysis techniques to investigate more autism subtypes.

Our preliminary investigation demonstrates that Big Data has the power to conduct genome-scale data-driven studies. However, it also has the flexibility to accommodate hypotheses-driven studies. We plan to use SNP combinations to connect autism-related genes from animal models to genes previously unaffiliated with autism.

Enabling Privacy-Preserving Biomedical Data Analytics in the Cloud and Across Institutions

Haixu Tang, Indiana University, Bloomington; *Xiaofeng Wang, Indiana University, Bloomington; Yan Huang, Indiana University, Bloomington; Xiaoqian Jiang, University of California, San Diego; Shuang Wang, University of California, San Diego; Lucila Ohno-Machado, University of California, San Diego*

The outsourcing of biomedical data into public cloud computing settings raises concerns over privacy and security. With respect to public cloud environments, there are concerns about the inadvertent exposure of human genomic data to unauthorized users. In analyses involving multiple institutions, there is additional concern about data being used beyond agreed research scope and being processed in untrusted computational environments. Significant advancements in secure computation methods have emerged over the past several years. A key challenge here, however, is how the analytical algorithms can be conducted in an efficient and cost-effective manner. Some of these techniques were implemented in general-purpose packages, they are not optimized to the analysis of biomedical data (note that these cryptographic algorithms are often computationally intensive, and thus the optimization to specific computational tasks can lead to acceleration of several magnitudes), and their applications are not straightforward. In our project, we attempt to develop a suite of methods and open source software tools that can be used by biomedical researchers in a plug-and-play manner for the statistical analysis of encrypted biomedical data. Our methods assume biomedical data will be protected by encryption after they are generated, and the subsequent analysis and sharing will always be performed on the encrypted form. We will utilize existing general-purpose encryption software, and will develop and optimize them for biomedical computation tasks. We will also evaluate these methods rigorously for their ability to support the analysis of various type of biomedical data. In the past two years, we organized twice the Critical Assessment of Data Privacy and Protection competition to assess the capacity of cryptographic technologies for secure human genomic computation in the cloud and cross-institutional collaborations. We will report our findings from the competitions and our future plan.

Multitask Deep Neural Net Kinase Activity Profiler

John Turner, University of Miami; *Bryce Allen, University of Miami; Stephan Schürer, University of Miami*

Deep neural networks (deep learning) is a powerful machine learning technology that has shown impressive success across a wide range of domains, including, recently, in drug discovery, utilizing structural features of small molecules to predict biological activity. Here we introduce the application of multitask deep learning to predict the activity of small molecules across the majority of the human Kinome. Small-molecule kinase inhibitors are an important class of anti-cancer agents and have demonstrated promising clinical efficacy in several different diseases; however resistance is often observed. Resistance mechanisms necessitate targeting multiple kinases or other targets in combination. Therefore, the prioritization of chemotypes with likely activity against one or more kinases can provide new therapeutic strategies for patients with resistant disease.

Our deep learning multi-task kinase predictors were built based on over 650K aggregated bioactivity annotations for over 300K diverse small-molecules covering 342 kinases curated from several sources. We demonstrated that our models perform significantly better than “classical” single task machine learning algorithms and we propose their potential for prioritizing desirable polypharmcology patterns.

Utilizing resources from the Library of Integrated Network-based Cellular Signatures (LINCS), we show that our predicted kinase activity signatures are highly concordant with compounds tested in the KINOMEscan assay.

In a proof of concept, we have implemented the technology as an app to enable researchers to generate a predicted kinase activity profile for any small molecule using a simple graphical user interface. Our goal is to develop this prototype into a robust public Kinome Profile Predictor for the benefit of the global research community.

Patient Linkage Across Research Datasets in a Patient Information Commons

Griffin Weber, Harvard Medical School; *Denis Agniel, Harvard Medical School; Tianxi Cai, Harvard T.H. Chan School of Public Health; Boris Hejblum, Harvard T.H. Chan School of Public Health; Isaac Kohane, Harvard Medical School; Shawn Murphy, Massachusetts General Hospital; PIC-SURE Patient Linkage Working Group, Harvard Medical School*

Patient-centered Information Commons: Standardized Unification of Research Elements" (PIC-SURE) is a BD2K Center of Excellence building a large, multi-center Patient-centered Information Commons (PIC), which can aggregate data about an individual patient that are scattered across disparate data sources. A key component of this is the ability to match records about the same patient that are in different datasets. We previously formed a Patient Linkage Working Group to address this challenge. Progress has been made on three fronts: (1) creating new linkage algorithms, (2) evaluating the performance of the algorithms on real datasets, and (3) implementing a patient linkage workflow within the PIC-SURE software architecture. Numerous methods have been described for linking patient records using direct identifiers, such as name, date of birth, and social security number. However, research datasets often remove these identifiers to protect patient privacy. We developed a probabilistic linkage algorithm to see if we can match patients based only on lists of diagnoses. Using a combination of electronic health records, administrative claims, and simulated data, we show that it is possible to accurately link patient records using diagnosis lists, but only when there is a high degree of similarity between the two datasets. To implement patient linkage in PIC-SURE, we created a Master Patient Index (MPI) to store linkage information across sites in the network. We did this by adding a "linkage probability" attribute to an existing patient mapping feature in the i2b2 (Informatics for Integrating Biology & the Bedside) software platform, in order to keep track of uncertainty in the patient matches. For traditional linkage using direct identifiers, we compared several products and found that the CDC's Link Plus program had the most complete set of features and scaled well to large datasets.

Visualizing Healthcare System Dynamics in Biomedical Big Data

Griffin Weber, Harvard Medical School; *Nicholas Benik, Harvard Medical School; Katy Borner, Indiana University; Nicholas Brown, Harvard Medical School; Daniel Halsey, Indiana University; Isaac Kohane, Harvard Medical School; Daniel O'Donnell, Indiana University*

Data in observational databases, such as electronic health records and administrative claims records, reflect both the pathophysiology (PP) of patients as well as their interactions with the healthcare system, which we call Healthcare System Dynamics (HSD). For example, the result of a laboratory test, such as a white blood cell (WBC) count, is a measure of the patient; but, the test was performed at that particular date and time because the patient visited her clinician, and the clinician determined that it was necessary to order the test. Investigators typically focus only on the test result, but the HSD component of the data often has greater predictive power. For example, counter intuitively, patients with an abnormal WBC test result at 3pm have higher three-year survival rates than patients with a normal WBC test result at 3am. This is because clinicians usually do not order laboratory tests at 3am unless they think the patient is very sick. The time of day of a clinical encounter, the day of the week, the amount of time since the last encounter, and total amount of data that has been recorded about a patient (the patient's "fact count") are just a few of the HSD concepts that are important in predicting different types of patient outcomes. In this BD2K Targeted Software Development project, we created an ontology and data visualization plugins for a widely used clinical research platform, i2b2, to help investigators learn about HSD and incorporate it into their studies. More broadly, this project illustrates, though data visualizations, how understanding the processes that lead to the generation of biomedical big data can be just as important as the data themselves.

KEGGlincs Design and Application: An R Package for Exploring Relationships in Biological Pathways

Shana White, University of Cincinnati; *Mario Medvedovic, University of Cincinnati*

The Library of Integrated Network-based Signals (LINCS) project is a data generation venture that is a quintessential example of current efforts concerning 'big data' in the biomedical research environment. One element of this project is the production of gene expression profiles corresponding to individual gene knock-outs within specific cancer cell lines. The R package 'KEGGlincs' and the companion data package 'KOdata', both published with the latest version of Bioconductor (3.4), were developed to promote synergy between existing pathway structures from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and LINCS data in order to reveal mechanisms of biochemical signaling processes that display heterogeneity across different types of cells.

KEGG pathways are manually-curated biological pathways represented as networks of nodes (genes) and directed edges whereby experimental evidence determines the nature and direction of an edge (relationship) between genes. The network structure for pathways that KEGG provides is a promising tool for bioinformatics research, and indeed there are existing methods for quantifying the level of pathway perturbation [within an experiment] that make use of KEGG pathways. However, existing approaches consider changes of gene expression only of the genes in a particular pathway and not changes in expression of downstream targets. This restricts the definition of perturbation to mean change in gene expression rather than a much broader, and perhaps more meaningful, change in gene function.

The LINCS data resource (KOdata) combined with the functionality offered by KEGGlincs allow for the investigation of relationships between genes in a given pathway in a cell-type-specific manner via analysis of overlapping de-regulated genes corresponding to pairs of experimental knockouts. This approach to pathway analysis yields quantitative measures and a novel method for annotating relationships (edges) between genes programmatically created in R and automatically visualized in an interactive session via Cytoscape software.

CEDAR: Easing Authoring of Metadata to Make Biomedical Datasets More Findable and Reusable

Debra Willrett, Stanford Medicine, Center for Biomedical Informatics Research; *John Campbell, Northrop Grumman, Inc.; Kei-Hoi Cheung, Yale University; Michel Dumontier, Stanford University; Kim A. Durante, Stanford University; Attila L. Egyedi, Stanford University; Olivier Gevaert, Stanford University; Rafael Gonçalves, Stanford University; Alejandra Gonzales-Bertran, Oxford University; John Graybeal, Stanford University; Purvesh Khatri, Stanford University; Steven H. Kleinstein, Yale University; Csongor I. Nyulas, Stanford University; Maryam Panahiazar, Stanford University; Philippe Rocca-Serra, Oxford University; Marcos Martínez-Romero, Stanford University; Susanna-Assunta Sansone, Oxford University; Ravi D. Shankar, Stanford University; Mark A. Musen, Stanford University*

Online biomedical repositories contain a wealth of freely available data submitted by the research community, but to reuse this data in further studies requires well-annotated associated metadata. There is a growing set of community-developed standards for creating this metadata, often in the form of templates; still, the difficulties of working with these standards are significant. The Center for Expanded Data Annotation and Retrieval (CEDAR) is building an end-to-end system to ease the authoring of metadata and the templates. This system targets the creation of higher quality metadata to facilitate data discovery, interoperability, and reuse. With our public release in September 2016, we now support many new features that make authoring easier.

Template and Metadata Repository: We developed a standardized representation of metadata and the templates that describe them, together with Web-based services to store, search, and share these resources. Templates created using CEDAR technology are stored in our openly accessible community repository, and can now be shared with other people and groups. Researchers can search for templates to annotate their studies, and share their metadata with others. We've now added Web-based interfaces and REST APIs to facilitate access to templates, and all the metadata collected using those templates.

Template and Metadata Editor: We developed highly interactive Web-based tools to simplify the process of authoring metadata and templates. The Template Editor allows users to create, search, and author templates. An upgraded feature provides interoperation with ontologies: interactive look-up services linked to NCBO's BioPortal (bioportal.bioontology.org) let template authors find ontology terms to annotate fields in their templates and to define possible values of fields, including creating new terms and value sets. The Metadata Editor, which creates a forms-based acquisition interface from a template, has been redesigned so users can more easily populate metadata based on the template fields.

BioThings APIs: Linked High-Performance APIs for Biological Entities

Chunlei Wu, The Scripps Research Institute; *Jiwen Xin, The Scripps Research Institute; Cyrus Afrasiabi, The Scripps Research Institute; Sebastien Lelong, The Scripps Research Institute; Ginger Tsueng, The Scripps Research Institute; Andrew I. Su, The Scripps Research Institute*

The accumulation of biological knowledge and the advance of web and cloud technology are growing in parallel. Recently, many biological data providers start to provide web-based APIs (Application Programming Interfaces) for accessing data in a simple and reliable manner, in addition to the traditional raw flat-file downloads. Web APIs provide many benefits over traditional file downloads. For instance, users can request specific data such as a list of genes of interest without having to download the entire dataset, thereby providing the latest data on demand and reducing computation and data transfer times. This means that programmers can spend less time on wrangling data, and more time on analysis and discovery.

Building and deploying scalable and high-performance web APIs requires sophisticated software engineering techniques. We previously developed high-performance and scalable web APIs for gene and genetic variant annotations, accessible at MyGene.info and MyVariant.info. These two services are a tangible implementation of our expertise and collectively serve over 4 million requests every month from thousands of unique users. Crucially, the underlying design and implementation of these systems are in fact not specific to genes or variants, but rather can be easily adapted to other biomedical data types such drugs, diseases, pathways, species, genomes, domains and interactions. We are currently expanding the scope of our platform to other biological entities. Collectively, we refer them as “BioThings APIs” (http://biothings.io).

We also applied JSON-LD (JSON for Linking Data) technology in the development of BioThings APIs. JSON-LD provides a standard way to add semantic context to the existing JSON data structure, for the purpose of enhancing the interoperability between APIs. We have demonstrated the applications of JSON-LD with BioThings APIs, including data discrepancy checks as well as the cross-linking between APIs.

Global Detection of Epistasis

Sihai Zhao, University of Illinois at Urbana-Champaign; *Congcong Chen, University of Illinois at Urbana-Champaign; Jieping Ye, University of Michigan*

The genetics of complex traits may be governed by intricate epistatic interactions, in addition to simple additive effects. Many existing methods for discovering epistasis are based on testing for statistical interaction between all possible pairs of typed variants, or some subset of these pairs. However, these approaches are extremely computationally intensive and tend to have low statistical power because of the extraordinary multiple testing burden. To avoid these issues, we test a slightly different question: whether a variant is involved in any pairwise interactions at all. Our global testing approach only requires testing all variants instead of all pairs of variants. Our test statistic is based on recently developed residual variance estimators for high-dimensional linear models, and we obtain p-values using a permutation procedure. We demonstrate the performance of our method in simulations and apply it to detect epistasis in Alzheimer's disease and other traits.

Curate Patient-Centric Multi-Omics Data for Precision Medicine

Jun Zhu, Icahn School of Medicine at Mount Sinai

A biological system is complex. Multiple levels of regulation enable it to respond to genetic, epigenetic, genomic, and environmental perturbations. To understand whether a patient has a risk for a specific disease or whether a tumor responds to a therapy, we need to examine the patient from genetics, epigenetics, genomics, proteomics, and metabolomics angles. Many predictive models have been generated to assess individual's risk or response based on each type of these omics data. Public databases, such as TCGA (The Cancer Genome Atlas), have been created for depositing diverse types of omics data for public dissemination. However, sample errors, such as sample-swapping or mis-labeling, sample cross-contamination, meta-data errors, are inevitable during the process of data generation and management. If a gene expression profile from a different patient was assign to our patient of interest, we may end up with complete wrong prediction and put out patient at a great risk for wrong medicine, which is totally opposite from the goal of precision medicine. Thus, to accumulate omics and clinical data for precision medicine, it is critical to properly curate patient-centric omics data so that different types of omics data pertaining to the same individual truly match to each other.

We developed multiple methods for finding intrinsic regulations between different omics data types, and developed barcodes to match different omics profiles to different people. We also developed methods for matching omics and meta-data, and methods for checking sample cross contamination based on intrinsic barcodes. We applied our methods into TCGA datasets and detected multiple sample errors in many cancer datasets in TCGA (such as glioblastoma, kidney, lung, prostate, stomach). These results suggest that sample errors are not a dataset specific problem but more global problem in public databases.

Collaborative Presentations

Aztec and CEDAR: Extraction of Digital Object Metadata From Free Text

Brian Bleakley, HeartBD2K Center at UCLA; *Chelsea Ju, HeartBD2K Center at UCLA; Vincent Kyi, HeartBD2K Center at UCLA; Justin Wood, HeartBD2K Center at UCLA; Patrick Tan, HeartBD2K Center at UCLA; Howard Choi, HeartBD2K Center at UCLA; Martin O'Connor, CEDAR Center at Stanford University; Wei Wang, HeartBD2K Center at UCLA; Mark Musen, CEDAR Center at Stanford University; Peipei Ping, HeartBD2K Center at UCLA*

Introduction

Software tools, datasets, and other digital objects are frequently described in free text publications that accompany their release. These descriptions are often unstructured and are not readily importable to existing repositories such as PRIDE for proteomics data or GEO for genomic data. The CEDAR suite of metadata tools provides a number of means to rapidly accelerate the production of structured metadata to adequately describe datasets in these repositories. Investigators who use CEDAR to generate and populate metadata templates can take advantage of time-saving predictive features such as auto-complete and pre-population of fields.

Methods and Results

Our collaborative effort builds upon the strength of CEDAR metadata suite and the Aztec platform; we aim to provide substantial enhancements to the CEDAR metadata suite by leveraging Aztec's text-mining system to identify datasets in biomedical journal publications, identify the template that most-closely matches the described experiment, and extract appropriate ontology terms from the text.

Using Stanford CoreNLP, we can identify terms in free text publications with specific semantic relationships to terms from the metadata template's associated ontologies. Additionally, the Grobid machine learning library provides a means to extract, structure, and analyze the descriptive data in these free text publications.

A RESTful, public-facing API will be released as part of the Aztec platform, which will enable other developers to leverage our metadata extraction system. An existing data repository could then significantly enrich its dataset's metadata by providing our API with a CEDAR template and associated publications from open access journals. This is an important step towards the widespread adoption of the FAIR Principals, ensuring that data is Findable, Accessible, Interoperable and Reusable.

Leveraging the CEDAR Workbench for Ontology-Linked Submission of AIRR Data to the NCBI-SRA

Syed Ahmad Chan Bukhari, Department of Pathology, Yale School of Medicine; *Kei-Hoi Cheung, Department of Emergency Medicine, Yale Center for Medical Informatics, Yale School of Medicine; Martin J. O'Connor, Center for Expanded Data Annotation and Retrieval, Stanford Center for Biomedical Informatics Research; John Graybeal, Center for Expanded Data Annotation and Retrieval, Stanford Center for Biomedical Informatics Research; Mark A. Musen, Center for Expanded Data Annotation and Retrieval, Stanford Center for Biomedical Informatics Research; Steven H. Kleinstein, Department of Pathology, Yale School of Medicine*

Next-generation sequencing technologies have led to a rapid production of high-throughput sequence data characterizing adaptive immune-receptor repertoires (AIRRs). As part of the AIRR community (http://airr.irmacs.sfu.ca) data standards working group, we have developed an initial set of metadata recommendations for publishing AIRR sequencing studies. These recommendations will be implemented in several public repositories, including the NCBI sequence read archive (SRA). Submissions to SRA typically use a flat-file template and include only a minimal amount of term validation. In order to ease the metadata authoring and to implement the ontological terms validation of repertoire sequence data, we are developing an interactive template through CEDAR workbench that will allow for ontological validation, and subsequent deposition in SRA. CEDAR workbench also allows the user to populate the template with metadata for data submission to various data repositories. The incorporation of template-element level ontology mapping not only facilitates validation of data submission, but also enables intelligent queries within and across repositories.

PREFIX: ACCESSION Compact Identifier Resolution: An EBI/CDL Collaboration

Tim Clark, Massachusetts General Hospital and Harvard Medical School; *Ian Fore, National Institutes of Health; Niall Beard, University of Manchester; Jeffrey Grethe, University of California, San Diego; Greg Janeé, California Digital Library; Rafael Jimenez, ELIXIR; Nick Juty, European Bioinformatics Institute; John Kunze, California Digital Library; Julie McMurry, University of California, Santa Barbara; Sarala Wimalaratne, European Bioinformatics Institute*

Many science policy studies insist that robust archiving and citation of primary research data should be a prerequisite for publication of any claims which rely upon such data. If these recommendations were to be implemented – along with parallel actions for research resources and software – we would have a far stronger assurance that published results will be reliable, and may be reused.

However, proper data citation requires that citation references are machine resolvable. That is, any reference should be appended with a globally unique, web-resolvable identifier. DOIs fit the bill here, but are used by a small minority of the over 500 specialized biomedical data repositories. The most common approach is to assign locally-unique accession numbers. How do we implement data and resource citation at scale without a massive re-engineering project across all relevant biomedical repositories?

How can locally assigned identifiers become globally resolvable?

We describe here a project to enable PREFIX: ACCESSION based identifier resolution using a common registry of resource namespace prefixes and provider codes. This common registry and resolution approaches have been agreed by two major identifier and metadata providers in Europe and the U.S.A., the Identifiers.org system of the European Bioinformatics Institute (EBI), and the Name-To-Thing resolver at the California Digital Library (CDL).

This is a subproject of the bioCADDIE Data Citation Implementation Pilot, organized as a FORCE11.org community group. We plan for both the EBI and CDL resolver systems to be capable of handling these coordinated compact identifiers by the end of 2016.

Cloud-Based Integration of Causal Modeling and Discovery Tools With a Unified Patient Research Database

Jeremy Espino, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; *Paul Avillach, Department of Biomedical Informatics, Harvard Medical School; Michael Davis, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Jeremy Easton-Marks, Department of Biomedical Informatics, Harvard Medical School; Michael McDuffie, Department of Biomedical Informatics, Harvard Medical School; David Bernick, Department of Biomedical Informatics, Harvard Medical School; Gregory Cooper, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh; Isaac Kohane, Department of Biomedical Informatics, Harvard Medical School; Michael Becich, Department of Biomedical Informatics, School of Medicine, University of Pittsburgh*

Both the NIH Commons and the BD2K initiatives share a common goal of providing a central computational environment for the sharing, reuse, interoperability, and discoverability of data. The University of Pittsburgh Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data (CCD) and the Harvard University Medical School (HMS) Patient-Centered Informatics Common: Standardized Unification of Research Elements (PIC-SURE) have implemented a framework that provides a model for biomedical researchers to share data and tools in a cloud environment. An instantiation of this framework allows researchers to apply causal discovery algorithms to biomedical and clinical big data to discover new and significant causal relationships. The PIC-SURE team has adapted their tranSMART tool and extended their application-programming interface (API) to provide secure access to the Autism Simons Simplex Collection dataset running on an Amazon Elastic Compute Cloud (Amazon EC2) instance. Through this service, a secure authentication layer, via Auth0, provides an institutional single sign-on which handles requests for access to the data. This data is then readily available for analysis by an Amazon EC2 implementation of the CCD causal web tool and API. CCD causal web provides a similar authentication method using institutional single sign-on for handling access control. Our proof-of-principle integrated data ecosystem serves as a model for sharing and analyzing biomedical big data in a secure and scalable manner in a cloud environment.

Annual California BD2K Centers Regional Meetings: Building Connections Across Centers

Ben Heavner, Institute for Systems Biology, The BDDS Team

Recognizing the geographic concentration of BD2K Centers in California, the BDDS and ENIGMA centers have hosted two annual "California Big Data Annual Meetings" for a highly interactive opportunity for BD2K centers to present software, big data discoveries, and big data resources and to discuss how to leverage our California-based BD2K efforts into further consortia activities for large-scale biomedical science and training. These meetings have included participants from both California (including centers affiliated with Stanford, UCLA, USC, UC San Diego, California State, Cal-Brain, and UCSC), as well as Washington, Illinois, and NIH program staff. The meetings have provided opportunities for each center to share their research, as well as to advance collaboration between California centers on tools, training and papers.

Participants in the 2015 meeting established working groups on "Reviewing Scientific Tools and Workflow Technologies", "Computing, Data, & Storage Infrastructures", "Establishing Educational and Training Experiences", and "Mathematics of Big Data Algorithms". The resulting intra-center conversations identified common challenges, technical approaches, and barriers to effective scientific progress using data at scale, along with possible avenues to collaboratively address these shared challenges. Specifically, collaborative groups were charged to advance 4 specific issues: "Registry and Catalogue", "Shareable Resources", "Sustainability", and "Quality Control".

In 2016, the meeting included opportunity for sharing of each Center's scientific results and demonstration of newly implemented technologies to facilitate bioscience research using big data, and discussion on how CA BD2K Centers can further work together on integration and use of tools. Prompted to consider what actions can we take together, participants identified 11 specific opportunities for intra-center data sharing and collaborative research opportunities. Already, preliminary efforts are under way to realize these opportunities, and the California BD2K centers are looking forward to another opportunity to meet together in 2017.

Machine Learning in Textual Data of Cardiovascular Disease via Phrase Mining and Network Embedding

Vincent Kyi, HeartBD2K Center at UCLA; *Doris Xin, KnowEnG Center; Leah Briscoe, HeartBD2K Center at UCLA; Fangbo Tao, KnowEnG Center; Yu Shi, KnowEnG Center; Brian Bleakley, HeartBD2K Center at UCLA; Jiawei Han, KnowEnG Center; Peipei Ping, HeartBD2K Center at UCLA*

The Medline Database comprises a treasure trove of over 2.2 million cardiovascular-related scientific articles, which are largely in unstructured data, making it a formidable challenge to identify datasets and to comprehend information. We aim to address this big data challenge by developing novel text-mining and machine-learning techniques to dissect textual data, to identify patterns in datasets, and to reveal mechanistic insights.

Textual datasets containing six cardiovascular diseases (CVD) were used as the validating corpus. We applied a combination of novel cutting-edge phrase-mining algorithms, namely SegPhrase+ and ToPMine, and a newly developed phrased-based networking embedding technique, Large-scale Information Network Embedding (LINE). From the main corpus, we evaluated 551,358 publications (dating from 1995 to 2016) in the Medline Database using Pubmed as a search engine. To extract articles and distinguish hidden patterns, we applied MeSH-terms from the National Library of Medicine's controlled vocabulary together with non-MeSH-term synonyms in CVD, in parallel with a list of top-250 proteins that are highly relevant in CVD. Each protein was scored using Context-aware Semantic Online Analytical Processing (CaseOLAP) and ranked within each CVD group. The final ranked score for each protein was calculated using a geometric mean of three criteria: 1) Integrity (meaningfulness and high-quality phrase), 2) Popularity (total count in the extracted articles of the CVD), and 3) Distinctiveness (count in the extracted articles of a CVD as compared to other 5 CVDs). The correlations of seed-pairs were computed by their coexistence frequency and their relationship uniqueness.

Our text-mining approach is the first study using CaseOLAP on textual data of cardiovascular health and disease, and highlights cardiovascular protein-disease relationships. We believe this approach affords promising biomedical and clinical applications including pattern discovery in accumulated patient information from electronic health records.

Revisions to the Disease Ontology to Support the Alliance of Genome Resources

Elvira Mitraka, Institute for Genome Sciences, University of Maryland School of Medicine*; Lynn M. Schriml, Institute for Genome Sciences, University of Maryland School of Medicine; Mary Shimoyama, Human and Molecular Genetics Center, Department of Surgery, Medical College of Wisconsin; Susan M. Bello, The Jackson Laboratory; Stanley J.F. Laulederkind, Human and Molecular Genetics Center, Department of Surgery, Medical College of Wisconsin; Cynthia L. Smith, The Jackson Laboratory; Janan T. Eppig, The Jackson Laboratory*

Model organism databases are one of the cornerstones of biomedical research, serving thousands of users daily. Each database curates and integrates vast amounts of genetic, functional, evolutionary, molecular, physiological and other biological data, information, and knowledge from the scientific literature, individual researchers, and a variety of publicly available sources. Six model organism databases and the GO consortium have formed a partnership to build the Alliance of Genome Resources (AGR), an integrated interspecies genome resource in support of translational research. The AGR will offer a unified resource to interrogate model organism data, including information related to human diseases, in order to advance genome biology and genomic medicine. To accomplish this, unifying data standards have been adopted for cross-model organism database use, including the use of the Disease Ontology (DO) as the standard vocabulary for human disease. Mouse Genome Database (MGD, http://www.informatics.jax.org) and the Rat Genome Database (RGD, http://rgd.mcw.edu), founding members of the AGR, are collaborating with DO in order to, on one hand, align their data with the disease classification model used by the DO, and, on the other other hand, to expand and enrich DO to include all the terms and relationships to cover their needs. As a result of this collaboration, the DO is now generating disease terms for individual members of OMIM's phenotypic series, has expanded the number of diseases with inferred anatomy based relationships, and has implemented a new relation to include database entries that show susceptibility to a particular disease. This will deeply enrich DO, while at the same time provide MGD and RGD - and the other members of the AGR - with a robust resource that will foster interoperability and provide the human genetics/genomics community with a consistent way to query disease associations.

When the World Beats a Path to Your Door: Collaboration in the Era of Big Data

Mark Musen, Stanford University

For many years, my laboratory has led major projects that provide computational infrastructure for work in data science. Protégé is a software system for editing ontologies that, over the past 20 years, has acquired more than 300,000 registered users, of whom nearly 200 have contributed more than 130 plug-ins for use by the community. The National Center for Biomedical Ontology (NCBO) provides BioPortal, a repository of most of the world's publicly available biomedical controlled terminologies and ontologies, as well as services that use those ontologies for a variety of tasks. More than 45,000 users access the NCBO ontology repository each month.

The Center for Expanded Data Annotation and Retrieval (CEDAR), supported by the BD2K initiative, is developing technology to assist scientists in the creation of comprehensive metadata to describe experimental datasets that will be stored in online repositories. CEDAR has already attracted the interest of many initiatives supported by the NIH, including HIPC, LINCS, HeartBD2K, and CaDSR. If CEDAR proves successful, it may likewise see many users and collaborations.

Experience with the Protégé and NCBO projects has taught us the challenges of interacting with large user communities and of supporting collaborations with a wide range of users. It is essential to have open communication channels, to develop software with collaboration in mind, and to manage collaborations actively to achieve maximum benefit. Effective collaboration in academic projects remains difficult, however, due to resource limitations—in particular, the availability of technically knowledgeable personnel who have the time to manage such collaborations and to help map out new directions and projects. I will discuss how the lessons of our experience with Protégé and NCBO are helping us to develop and enrich our collaborations on the CEDAR project, and how appropriate communication with users can help to scale these interactions.

ELIXIR: A European Distributed Infrastructure for Life-Science Information

***Pablo Roman-Garcia, ELIXIR**; Andrew Smith, ELIXIR Hub; Serena Scollen, ELIXIR Hub; Norman Morrison, ELIXIR Hub; Rita Hendricusdottir, ELIXIR UK - University of Edinburgh; Jo McEntyre, ELIXIR EBI – EMBL; Niklas Blomberg, ELIXIR Hub*

ELIXIR connects National Bioinformatics Institutes and EMBL-EBI into a distributed European infrastructure for biological research data. The structure is formed by 20 "Nodes" (National Bioinformatics Institutes) and a Hub (a central coordination Office). ELIXIR services, most of which are run from ELIXIR Nodes, include data resources (data access, archives), bioinformatics tools (Registry: bio.tools.org), compute provision, standards development and training.

Global Initiatives

Currently, EMBL-EBI, which is the largest "Node" in ELIXIR, receives over 22 million requests a day from global users. Human Protein Atlas (HPA) (ELIXIR Sweden) received more than 750,000 visits during 2013, of which approximately 60% was from outside of Europe.

From the International point of view, ELIXIR is identifying a set of Core Data Resources that are globally competitive and of critical importance to the life science community and will actively promote their integration and sustainability.

In addition, ELIXIR provides researchers all over the world with information about the bioinformatics training courses and materials available in Europe, and where they can be found. It does this through the Training eSupport System (TeSS) portal.

In the field of human genomics ELIXIR is establishing a sustainable and global infrastructure for Human Genomics and Translational Data, establishing solutions to support data access and sharing such as Beacons in the European Genome-phenome Archive and genomics resources in ELIXIR Nodes with individual-level data from biomedical research projects, supporting a three-level access system. This work is core in the Global Alliance for Genomics and Health (GA4GH)-ELIXIR collaboration.

BD2K Collaboration

ELIXIR has developed a range of collaborations and alignment with BD2K. A Collaboration Strategy for Training is in place, with joint workshops between ELIXIR and BD2K, for example in Research Object Identifier- ROI. The latter points to common efforts in interoperability, as ELIXIR and BD2K have similar aims, especially in Data FAIR-ification.

Count Everything: Secure Count Query Framework Across Big Data Centers

Ida Sim, University of California, San Francisco; *Xiaoqian Jiang, University of California, San Diego; Claudiu Farcas, University of California, San Diego; Kevin Osborn, University of California, Santa Cruz; David Steinberg, University of California, Santa Cruz; Benedict Paten, University of California, Santa Cruz; Mark Diekhans, University of California, Santa Cruz; Paul Avillach, Harvard University; Wallace Wadge, Overlap Health; David Haussler, University of California, Santa Cruz; Isaac Kohane, Harvard University; Lucila Ohno-Machado, University of California, San Diego*

Electronic health record (EHR), mobile health and genome sequence data for the same patient are often spread across multiple geographically distributed hosts that expose different application program interfaces (API). The lack of secure and privacy-preserving interfaces impedes the ability of applications to integrate heterogeneous data from different sources. Because all of statistics start with counts of observations, developing a federated count query infrastructure is a first step to supporting more advanced queries.

The Count Everything project supports count query interoperability across three BD2K Centers [1]: the Mobile Sensor Data to Knowledge (MD2K) Center houses mobile health data [2]; the Center for Big Data in Translational Genomics (CBDTG) [3] houses genomic sequence data; the Patient-centered Information Commons: Standardized Unification of Research Elements (PIC-SURE) Center [4] houses EHR data. The NIH National Center for Biomedical Computing Integrating Data for Analysis, Anonymization, and Sharing (iDASH) [5] and the BD2K Biomedical and healthcare Data Discovery Index Ecosystem (bioCADDIE) [6] -- also participate in this project. We developed a secure protocol to decompose a user query into subqueries and return the cardinality of matched items that share the same standardized patient IDs. The three data-holding BD2K Centers exposed data to the Count Everything infrastructure via their respective APIs.

We created a synthetic cohort by assigning unique global IDs to convenience sets of mobile, genomic 1000 Genomes, and clinical CDC NHANES data, which are queried by the Count Everything infrastructure. To safeguard the intermediary results and return only the final count, we developed a secure framework based on homomorphic encryption (HME) [7], which can support computation on encrypted data. This is a semi-centralized system, which utilizes iDASH as the HME processor and bioCADDIE as a cryptographic service provider (CSP) in a hub-and-spoke architecture. The web-based service is available for testing: http://counteverything.ucsd-dbmi.org:8080/JavaBridge/beta2/index.html#.

FAIR LINCS Data and Metadata Powered by the CEDAR Framework

Raymond Terryn, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; *Amar Koleti, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Rafael S. Gonçalves, Stanford Center for Biomedical Informatics Research; Vasileios Stathias, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Michele Forlin, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Wen Niu, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, BD2K LINCS Data Coordination and Integration Center; Caroline Monteiro, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, BD2K LINCS Data Coordination and Integration Center; Daniel Cooper, BD2K LINCS Data Coordination and Integration Center; Christopher Mader, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Vance Lemmon, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Dusica Vidovic, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; Csongor I. Nyulas, Stanford Center for Biomedical Informatics Research, BD2K CEDAR (Center for Expanded Data Annotation and Retrieval); Martin J. O'Connor, Stanford Center for Biomedical Informatics Research, BD2K CEDAR; Mario Medvedovic, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati, BD2K LINCS Data Coordination and Integration Center; Avi Ma'ayan, Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, BD2K LINCS Data Coordination and Integration Center; Caty Chung, Center for Computational Science and University of Miami, BD2K LINCS Data Coordination and Integration Center; John Graybeal, Stanford Center for Biomedical Informatics Research, BD2K CEDAR; Mark A. Musen, Stanford Center for Biomedical Informatics Research, BD2K CEDAR; Stephan Schürer, Center for Computational Science and University of Miami, Department of Molecular and Cellular Pharmacology, University of Miami, Division of Biostatistics and Bioinformatics, Environmental Health Department, University of Cincinnati*

The Library of Integrated Network-based Signatures (LINCS) program generates a wide variety of cell-based perturbation-response signatures using diverse assay technologies. For example, LINCS includes large-scale transcriptional profiling of genetic and small molecule perturbations, and various proteomics and imaging datasets. We have developed data processing pipelines, and supporting informatics infrastructure to access, standardize and harmonize, register and publish LINCS datasets and metadata from all Data and Signature Generating Centers (DSGC's). Metadata standards specifications provide a foundation for harmonizing and integrating LINCS data. Here we introduce a CEDAR-based LINCS Community Metadata Environment, to support end-to-end metadata management framework that supports authoring, curation, validation, management, and sharing of LINCS metadata, while building upon the existing LINCS metadata standards and data-release workflows. Following this initial validation, our goal is to create reusable metadata modules with user-friendly templates for each of the LINCS metadata categories and to make our suite of tools compatible with the CEDAR metadata technologies. This should further simplify metadata handling in the LINCS consortium and facilitate a global metadata repository at CEDAR. As other projects apply the same approach, many more datasets will become cross-searchable and can be linked optimizing the metadata pathway from submission to discovery.

Worldwide Big Data Collaborations: Examples From ENIGMA, Spanning 35 Countries

Paul Thompson, ENIGMA Center for Worldwide Medicine, Imaging and Genomics, University of Southern California

ENIGMA is a worldwide scientific alliance of 700 scientists – across 340 institutions from 35 countries – pooling their imaging, genetic, and clinical data to study 18 brain diseases (http://enigma.ini.usc.edu). By drawing on "big data", expertise, and computing infrastructure in 35 countries, the 30 working groups in ENIGMA have performed the largest neuroimaging studies to date of major depression, schizophrenia, bipolar disorder, and obsessive-compulsive disorder, mapping disease effects on the brain, and discovering protective or adverse factors influence these disease effects. We highlight 4 major areas where worldwide collaboration in big data has helped ENIGMA and brain science: (1) vastly increased statistical power to tackle new kinds of questions – such as ENIGMA's discovery of genetic loci that affect the brain; (2) better testing of the reproducibility and generalizability of findings, as in ENIGMA's maps of the major mental illnesses; (3) distributed computing, which speeds discovery by harnessing computer infrastructure internationally, (4) crowd-sourcing of ideas, where novel insights are made by mathematicians, geneticists and others not previously part of the Big Data community.

Sustainability

The Stewardship Gap

***George Alter, University of Michigan**; Fran Berman, Rensselaer Polytechnic Institute; Myron Gutmann, University of Colorado; Jeremy York, University of Michigan*

The Stewardship Gap Project is an 18 month effort to do a strategic comparative case study that sheds light on the size, characteristics and sustainability of valuable sponsored research data and creative work. To use data now and in the future requires that projects and organizations take responsibility for the stewardship (current management) and preservation (management over time) of data on which modern research and creative work depend. Yet even as the importance of research data increases, we know little about the quantity, characteristics or sustainability of those data. It is broadly suspected that there is a "stewardship gap" between the amount of valuable data developed as part of sponsored research/creative work or used in a research publication, and the amount of data that is at risk for loss or damage. This presentation will discuss interviews with scientists who were asked why their data have value and how they were prepared to preserve their data for the future.

The Role of Trustworthy Digital Repositories in Sustainability

David Giaretta, Giaretta Associates

Can all repositories be trusted to keep their holdings usable over the long-term? Will any of them be funded forever? Trustworthy Digital Repositories are part of a chain of preservation which begins with the creation of the data and proceeds to the handover to the next in the chain. A necessary but not sufficient condition is that the appropriate “metadata” be collected at each stage.

How can repositories be independently evaluated, what is the role preservation plays in interoperability and sustainability, and what metadata must be collected at which stage?

In brief, preservation requires ensuring continued usability/ understandability, first by those intimately connected to the creation of the data and eventually probably by those who know nothing about its origin or even the specific discipline. How can this be achieved?

This presentation outlines the concepts and processes which are needed to ensure this is possible and how to judge whether it is being done correctly. It will briefly describe how to justify preservation, how to help with interoperability here and now – as an integral part of the chain of preservation – and how to be confident that a repository can play its part in preserving the information.

Archiving Interpretations of Variants in ClinVar

MelissaLandrum, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

ClinVar is NCBI's archive of human genetic variants and interpretations of their relationship to disease. As an NCBI database, its goal is to gather this data on a national and international scale; provide long-term storage of the data; and facilitate use of the data by the research and medical communities. This presentation will discuss decisions surrounding data to include in ClinVar and how community input is gathered.

Combining Protein and Genome Annotation for Interpretation of Genomic Variants

Peter McGarvey, Georgetown University Medical Center; *Andrew Nightingale, European Bioinformatics Institute; Hongzhan Huang, Department of Computer and Information Science, University of Delaware; Maria Martin, European Bioinformatics Institute; Cathy Wu, Department of Computer and Information Science, University of Delaware; Subha Madhavan, Innovation Center for Biomedical Informatics, Georgetown University Medical Center; Alex Bateman, European Bioinformatics Institute*

Genomic variants cause deleterious effects through many mechanisms, including affecting gene transcription and splicing, altering translation and disrupting key components of a protein's structure and function. Understanding the effects of non-synonymous single nucleotide polymorphisms (SNPs) on protein function is a key component for gene and variant curation, but this information is not readily available at the genome level. There is a need for improved synchronization between the genome and the proteome resources to be able to map from the variant in the DNA sequence to that in the protein sequence and to the actual biological consequence of the variant. UniProtKB contains decades of effort in describing protein functions and features including amino acid variation via literature-based and semi-automated expert curation. To integrate UniProtKB, information needs to be mapped from protein coordinate space onto the coordinate space of genomic sequences. UniProt has now mapped protein feature annotation in the human proteome to the GRCh38 build of the human genome. Twenty-six structural and functional features plus isoforms are provided including: enzyme active sites; modified residues; protein binding domains; protein isoforms; and protein variations. The mappings and related annotation are available as public genome tracks on the UCSC and Ensembl genome browsers and programmatically via a data API and as text. UniProt feature and variation data is being loaded into the ClinGen Knowledge Base and annotated protein variations are loaded as evidence documents into the ClinGen Pathogenicity Calculator. Examples where genomic variants are aligned with UniProt features such as Active Sites, Domains and other features that could explain a variant's pathogenicity will be presented as well as a comparison of UniProt disease associated amino acid variations with ClinVar pathogenic SNPs. The results suggest that this integration can help with the classification of pathogenic variants and improve interoperability between the proteomics and genomics communities.

Interoperability of NURSA, PharmGKB, dkNET, and DataMed

Neil McKenna, Baylor College of Medicine; *Yolanda F. Darlington, Duncan Comprehensive Cancer Center Biomedical Informatics Group, Baylor College of Medicine; Ryan Whaley, Pharmacogenomics KnowledgeBase, Departments of Genetics and Bioengineering, Stanford University School Medicine; Alexey Naumov, Duncan Comprehensive Cancer Center Biomedical Informatics Group, Baylor College of Medicine; Wasula Kankanamge, Duncan Comprehensive Cancer Center Biomedical Informatics Group (YFD, AN, WK), Baylor College of Medicine; Teri E. Klein, Pharmacogenomics KnowledgeBase, Departments of Genetics and Bioengineering, Stanford University School Medicine; Michele Whirl-Carillo, Pharmacogenomics KnowledgeBase, Departments of Genetics and Bioengineering, Stanford University School Medicine; Jeffrey S. Grethe, Center for Research in Biological Systems, University of California, San Diego*

Nuclear receptor signaling pathways are of considerable interest in the development of sensitive and specific therapeutics for the treatment of metabolic disease and cancer. The initial and current funding periods of the One of the goals of the BD2K initiative is to mandate the scalable, low cost interoperability of different data resources to make individual datasets more widely available to the research community. The Nuclear Receptor Signaling Atlas (NURSA) is developing a large-scale transcriptomic resource that documents the tissue specific regulation of gene expression by NR signaling pathways, direct access to which would benefit the progress of research in numerous distinct subdisciplines. The NIH funds informatics resources with scientifically related but administratively distinct research constituencies, such as the effect of genetic variation on drug action in humans (the Pharmacogenomic Knowledge Base), and the regulation of metabolism and its derangement in metabolic disease states (the National Institute of Diabetes, Digestive and Kidney Disease Research Network (dKNET)). Given the cross-cutting relevance and potential utility of the Transcriptomine dataset to both resources, we are setting out in this proposal to develop a pipeline that will provide for exposure of Transcriptomine data points via application programming interfaces (APIs), and to collaborate with the PharmGKB and dKNET resources to demonstrate the use of these APIs to make the data points available to users of these resources. In turn NURSA will consume data points and associated metadata elements from PharmGKB and DKNET of relevance to NURSA's userbase and expose these on the NURSA website. Finally, NURSA datasets are exposed to the broader research community via the DataMed data discovery index. The multiple points of access created by NURSA's interoperability with these and other key nodes in the digital biomedical research ecosystem will drive the re-use of valuable discovery scale assets for the entire research community.

Interoperability, Sustainability, and Impact: A UniProt Case Study

Cathy Wu, University of Delaware

This presentation will use UniProt as a use case to discuss the issues of interoperability and sustainability, as well as metrics to measure impact. Topics will include (i) practices of UniProt as a central hub of biological resources that promotes FAIR principles, (ii) an analysis on sustainability of literature curation, balancing the needs for automatic annotation and expert curation for scientific quality and scalability as a reference database, and (iii) metrics to measure quality, utilization and impact of the resource.