2017 Biomedical Computation at Stanford Symposium

Abstract Book

Stanford University

April 10, 2017

Contents

Keynote Talk Titles

Using Big Data to Manage Health and Disease

Michael Snyder, PhD
Stanford University

Estimating disease heritability using 13 million patients records

Nicholas Tatonetti, PhD
Columbia University

Statistical methods for quantitative proteomics

Olga Vitek, MS, PhD
Northeastern University

Discovering how drugs and their targets work through simulation and machine learning

Ron Dror, MPhil, PhD
Stanford University

Student Talk Abstracts

Deep learning approaches for functional variant prioritization

Anna Shcherbina
Stanford University

The project will develop a deep learning algorithm that assigns context-specific local functional scores to coding and non-coding variants throughout the genome, focusing on accurate and specific detection of multiple types of SNPs (rare germline, common germline, somatic, and others). The convolutional neural network will be trained on public datasets including Clinvar, RoadMap, and ENCODE, and will be applied to detect sub-threshold functional variants in GWAS studies and to identify causal variants in cardiovascular disease patients of the Stanford Clinical Genome Service. Finally, effects of gene-by-environment interactions on variant function will be investigated by adding accelerometry data to the model.

Iris: A Natural Language Platform for Biomedical Data Analysis

Ethan Fast
Stanford University

Modern high-throughput technologies such as RNA-seq are providing biologists with increasingly large amounts of high-dimensional data. Unfortunately, many biologists lack the programming skills necessary to mine statistical patterns or build predictive models from such data. These scientists would benefit enormously from a data scientist trained in biological terminology who could translate their ideas into executable code. Our work explores the question: could that data scientist be a machine?

We present Iris, a conversational artificial intelligence (AI) designed to help non-experts with data science tasks. Iris interacts with users through natural language to connect their words with executable code. For example, you can ask Iris to run a t-test on the control and experimental groups or find the most important features to predict cancer type. To process a user request, Iris first connects it with an executable command (e.g., a command for a two-sided t-test) and then converses with the user to resolve that commands arguments (e.g., control and experimental groups). Users can combine commands through composition (calling one command within another) and sequencing (calling a command on the result of a previous command) to achieve complex goals. For example, you might tell Iris to run a one-sample t-test against a mean of 3. And Iris would ask, Sure, what population should I use? And you might say, Use the control_group column from data.csv (a form of composition, executing one command inside another). And Iris would respond, Okay. Comparing that to a mean of 3, I got a p-value of .2. You might then ask, Is that statistically significant? (a form of sequencing, using the previous conversation in a new request). And Iris would say, No, the p-value is not less than 0.05.

Iris can also help non-experts apply best practices for machine learning. For example, suppose that Iris has created a Random Forest model called cancer-model to predict cancer subtypes. You tell Iris, Help me validate the cancer-model. Iris asks, Do you have any held out data? And you say, No. So Iris says, Sure, lets do cross-validation.

Today, Iris is in small-scale deployment across groups in Stanfords computer science and genetics departments. The platform is based on Python and supports natural language bindings for common

mathematical functions and machine learning models. In the near term, we are working with biologists and bioinformaticians to improve the systems interface and expand its set of functions. We are planning for a public release of Iris by the summer of 2017.

Proteogenomic analysis of surgically resected lung adenocarcinoma

Michael Sharpnack, Nilini Ranbaduge*, Arunima Srivastava*, Ferdinando Cerciello, Simona Codreanu, Daniel Liebler, Wayne Miles, Robert Morris, Jason McDermott, James Sharpnack, Joseph Amann, Chris Maher, Raghu Machiraju, Vicki Wysocki, Ramaswami Govindan, Parag Mallick, Kevin Coombes, Kun Huangˆ, and David Carboneˆ. *These authors contributed equally to this project, ˆCo-Principle Investigators.

Ohio State University

Despite apparently complete surgical resection, approximately half of resected early stage lung cancer patients relapse and die of their disease. Adjuvant chemotherapy reduces this risk by only 5-8%. Thus there is a need for better identifying who benefits from adjuvant therapy, the drivers of relapse and novel targets in this setting. We present a rationale and framework for the incorporation of high-content RNA and protein measurements into integrative biomarkers and demonstrate the potential of this approach for predicting risk of recurrence in a group of lung adenocarcinomas. In addition, we characterize the relationship between mRNA and protein measurements in lung adenocarcinoma and show that it is outcome specific. Our results suggest that mRNA and protein data possess independent biological and clinical importance, which can be leveraged to create higher-powered expression biomarkers

---

Deep learning approaches for functional variant prioritization

Peyton Greenside

Stanford University

Genomics-specific deep learning architectures reveal mechanisms of gene regulation”,”Transcription factors (TFs) bind combinatorial DNA sequence patterns in non-coding genomic elements to regulate chromatin and transcriptional programs that determine cell identity. These programs generate dynamic epigenomic profiles across diverse cell types in the human body. Deep learning models have achieved state-of-the-art predictions of epigenomic profiles from DNA sequence in an effort to uncover the regulatory mechanisms guiding each unique cell type. However, deep learning architectures used thus far in genomics are often directly ported from computer vision and natural language processing applications with few, if any, domain-specific modifications. We developed four new convolutional neural network layers that leverage the reverse-complement property of genomic DNA sequence by sharing parameters between forward and reverse-complement representations in the model to guarantee that both sequences produce identical predictions. We combine genomics-specific architectures with the power of interpretability using DeepLIFT and the ability to learn de novo context-specific motifs using MoDISco to understand mechanisms of gene regulation. We apply these techniques to investigate transcription factor binding and to understanding the regulatory programs guiding the differentiation of hematopoietic stem cells into diverse blood cell types.

Pathway and mechanism of antagonist binding to opioid receptors

Robin Betz
Stanford University

G protein-coupled receptors (GPCRs) are an important class of signaling protein of special interest for pharmaceutical development. Opioid receptors are a notable subset of GPCRs that are of interest in pharmaceutical development for new painkillers, ideally without the potential for addiction of current opioid drugs. Structure-based drug design is challenging at these receptors due to the difficulty of obtaining GPCR crystal structures. Subtle differences in molecules on the same scaffold can result in the molecule behaving as an agonist or antagonist, and minor alterations can produce large changes in selectivity among receptor subtypes.

Using a novel adaptive sampling method, we depart from the more traditional analysis of already bound ligands to computationally explore pathways and mechanisms of ligand binding in an unbiased manner that requires no prior knowledge of bound pose. A large number of short MD simulations are run in parallel, each exploring different regions of possible protein and ligand conformations. After each generation of simulations is complete, machine learning methods are used to identify those in which the ligand progresses along a possible binding pathway. No knowledge of the pathway or bound pose is required, making the method applicable to ligands where no similar crystal structure exists as well as to the identification of cryptic binding pockets and allosteric sites.

We apply the method to several opioid antagonists across with the goal of identifying differences in binding pathway across receptors that may in part be responsible for ligand functional selectivity. We find that the binding pathways of naloxone, naltrindole, and a multifunctional peptide at the mu- and delta-opioid receptors share a common entry pathway between transmembrane helices (TMs) 1 and 2 before proceeding deeper into the binding pocket. This is surprisingly different from the binding pathways of other GPCR ligands, such as dihydroalprenolol at the beta 2 adrenergic receptor, which enters between TMs 5 and 6.

Computationally determining time-resolved ligand binding pathways at all-atom resolution provides new insight into opioid antagonist selectivity and binding kinetics.

Poster Abstracts

A robust host-based gene expression diagnostic for malaria versus other infectious diseases

Aditya Rao, TE Sweeney, P Khatri
Stanford University

The global incidence of malarial infection is approximately 200 million, with annual mortality of at least 438,000. While early diagnosis of malaria can lead to rapid treatment and cure, the similarity of symptoms to other infectious diseases and the long incubation time causes delayed and incorrect diagnoses. The gold-standard diagnostic, blood smear, is time-consuming and not widely available. Rapid molecular tests for malarial parasites suffer from inaccuracy and relatively poor sensitivity. We have previously shown that diagnostics based on host gene expression can be an effective way to diagnose infectious diseases such as tuberculosis, sepsis, and bacterial infections. Here we hypothesized that transcriptomic profiles of patients infected with malaria would yield a robust diagnostic for malaria versus other infectious diseases.

We performed a systematic search for gene expression datasets profiling malaria or other common tropical infectious diseases (dengue, typhoid, and other parasitic infections) using either PBMCs or whole blood. We used our previously-described COCONUT co-normalization platform to pool gene expression data from all 44 cohorts into a single group. This group was randomly split into 70% / 30% training/validation groups. We then used machine learning to derive a set of genes optimized for diagnostic power in the training group, which we then tested in the held-out validation group as well as an independent validation dataset.

We here show that a small set of host genes can be used to form a robust classifier for malaria vs. other infectious diseases that are within the differential diagnosis for malarial symptoms. As has been shown with the GeneXpert system (among others), gene-expression based diagnostics can be made to be rapid and affordable enough to become useful in low- and middle-income countries. Our 9-gene and 4-gene classifiers will need prospective validation prior to clinical translation.

Dimensionality Reduction of Neural Dynamics Within and Across Trials By Tensor Decomposition

Alex Williams
Stanford University

Decision-making, sensation, and motor behaviors occur within fractions of seconds, while memories and learned behaviors can require many days or months to mature. Recent advances enable long-term experiments that monitor all of these timescales across hundreds of neurons and behavioral trials. However, classic and commonly used dimensionality reduction techniques are ill-equipped to summarize data across multiple timescales. We represent multi-trial neural data as a tensor (i.e., a higher-order array), and apply canonical polyadic (CP) tensor decomposition to identify low-dimensional factors that separately summarize short-term and long-term changes in neural population dynamics. In synthetic data generated from model networks, this approach precisely identifies network inputs that vary across trials, whereas classical methods (PCA and ICA) fail to recover these signals. In experimental datasets collected from different species, brain regions, and behavioral tasks, CP decomposition uncovers sub-populations with interpretable within-trial as

well as across trial dynamics, reflecting behavioral strategies, experimental structure, rewards, and perturbations to the behavioral task. Specifically, we validate our approach (a) on neural activity in motor cortex in a primate executing repeated reaches with a virtual cursor controlled through a brain-machine interface (BMI), and (b) on prefrontal cortex activity in mice executing rule-based navigational strategies in a four-armed mazed. In both cases, perturbations to the experimental task caused changes in trial-to-trial neural dynamics, which were identified and compactly summarized by CP decomposition. We also develop specialized methods for fitting CP decompositions to spiking neural data by adding smoothness and nonnegativity constraints on the latent factors, and we draw novel connections between CP decomposition and existing neuroscience literature on gain modulation and trial-to-trial variability.

Profiling immune system sex differences in the healthy human transcriptome

Erika Bongen
Stanford University

Women are at higher risk of autoimmunity, while men are more likely to die of infectious disease. The molecular factors driving this phenomenon may be detectable in the transcriptome, as it is regulated by sex chromosomes and hormones. Previous studies on blood transcriptome sex differences were limited by studying a single population. No studies have systematically examined sex differences within the context of the immune system across multiple independent cohorts. We performed an immunologically focused investigation of robust transcriptional sex differences across global populations. First, we performed an integrated multi-cohort analysis of 6 cohorts consisting of 458 samples to identify a 178-gene signature, called the Immune Sex Expression Signature (iSEXS), which is differentially expressed between healthy male and female adults in the blood across populations. We validated iSEXS in 3 additional cohorts of 128 samples. Second, we examined sex differences in immune cell frequencies to determine whether iSEXS was driven by cell frequencies or phenotype. Using deconvolution, a method of predicting cell frequencies from bulk gene expression, we performed a meta-analysis sex differences in cell frequencies across populations. We validated our results in an independent mass cytometry dataset and found that females had higher levels of CD4+ T cells while males had higher levels of monocytes. Third, we examined the role of sex hormones and chromosomes in the regulation of iSEXS. We observed that 25% of iSEXS is located on the sex chromosomes. Importantly, in a cohort of disorders of sexual development, XY-individuals with normal female genitalia expressed the iSEXS at similar levels as XY-males, indicating that the iSEXS is primarily driven by chromosomal differences. As a robust gene signature across populations, iSEXS has applications in understanding why men and women have differential risks of autoimmunity and infection.

Mathematical Model of Cancer Heterogeneity & Biomarker Shedding Kinetics

Gautam Machiraju
Stanford University

Ovarian Cancer is the deadliest of gynecologic cancers with a 46.2% survival rate over 5 years largely in part to a lack of early detection methods. While known biomarkers exist for ovarian cancers, the most well-characterized, CA125, is used in screening tests due to its known properties, but also lacks specificity in screening asymptomatic patients. This underscores the importance of early detection methods that find a correlation between the concentration of a target biomarker and tumor size.

Early cancer detection plays a crucial role in prognosis and treatment. Mathematically, it has been shown that compartment models can be employed for understanding how detectable biomarkers are shed into the bloodstream. We propose a generalized model that attempts to

simulate a tumor cross-section and its corresponding cancer cell heterogeneity with the use of collected biomarker data as parameters for our model. The goal is to the correlate the concentration of a marker in blood as a proxy for tumor size, thereby allowing us to better infer the cancers stage. With the employment of numerical methods, preliminary results show both simulations of tumor growth, as well as biomarker shedding kinetics given a vascularized ovarian tissue microenvironment and target biomarker CA125.

Global Biobank Engine: An Online Tool for the Statistical Exploration of Large Genomic Datasets

Greg McInnes

Stanford University

Recent advances in sequencing and genotyping technologies have led to the development of many large genetic and clinical data repositories called biobanks. These biobanks harbor valuable data that may lead to a plethora of scientific discoveries, but they are often siloed within institutions and unavailable to the scientific community. In an effort to facilitate public use and exploration of the genetic data we have developed a website, the Global Biobank Engine (GBE), that allows users to perform statistical analysis of the genetic data as it relates to the clinical labels from individual biobanks. Statistical methods available through the browser include genome-wide association, phenome-wide association, and MRP (a Bayesian framework for assessing multiple rare variants and phenotypes). The initial version of GBE allows for public exploration of the UK Biobank data. Here we present various workflows as a demonstration of how to use the browser, as well as results from a specific investigation into asthma discovered using GBE.

Meta Analysis of Microbiome Data and Electronic Medical Data Provides New Insights On Preterm Births

Idit Kosti

University of California, San Francisco

Preterm birth is defined as the birth of an infant before 37 weeks of gestational age. It is the leading cause of perinatal morbidity and mortality worldwide, with 15 million preterm births per year. The human microbiome plays a crucial role in disease and a healthy state. It is known to transform throughout pregnancy, and can even affect the fetus before delivery. Studies on pregnancy microbiome are usually conducted on small cohorts due to the difficulty of recruitment and cost of sampling. In this study we integrated microbiota raw sequences from 5 pregnancy related studies ranging over 300 pregnant patients with more than 4,000 samples over multiple body sites. Due to the larger sample size and greater statistical power, computational meta analysis of proprietary and publicly available studies can provide us with the opportunity to gain new insights. Our results show variations in the microbiome between term and preterm pregnancies, and reveal new strains that might be involved in preterm birth. Using electronic medical data we were able to show preterm birth patients have different clinical features, such as allergies, compared to normal pregnancies. New hypothesis emerging from such an integrative analysis can lead to new possible treatment that will alter the microbiota and possibly prevent early delivery.

Characterization of Hypertrophic Cardiomyopathy using Wearable Sensors

Jessica Torres

Stanford University

Cardiovascular disease is the leading cause of death in the United States. The enormous burden posed to public health by cardiovascular disease can be reduced with the help of proper monitoring, diet, and exercise. Currently, within our healthcare system, a patients medical status is tracked via snapshots from clinical visits, which is suboptimal when some cardiovascular disease symptoms are asymptomatic and latent. This presents an opportunity for simple methods for tracking cardiovascular health over time outside of the hospital setting for patients with long-term disease associated instability in cardiovascular function. In this context, recent advances in wearable sensors can now provide the possibility of near-continuous monitoring for an exhaustive perspective of an individuals health. Specifically, a non-invasive sensor technology (PPG) holds the ability to measure physiological variables in an inexpensive, and easy to use manner. These signals captured by wearable sensors contain the opportunity to describe aspects of an individuals health status that can, in turn, be leveraged for prediction and classification of possible risk factors associated with disease. Here, I will be discussing my method for assessment of cardiovascular function using data from wearable sensors to characterize hypertrophic cardiomyopathy patients.

Compression of genomic sequencing reads with and without preserving the order

Shubham Chandak, Kedar Tatwawadi

Stanford University

Motivation: New Generation Sequencing (NGS) technologies for genome sequencing produce large amounts of short genomic reads per experiment, which is highly redundant and compressible. However, general-purpose compressors are unable to exploit this redundancy, due to the special structure present in the data. Although numerous specialized FASTQ compressors have been presented in the past few years, there has been no systematic analysis of the read compression problem and its limits in the literature.

Results: In this work, we first analyze the limits of read compression by computing bounds on the entropy of the reads. We then analyze the performance of some of the state-of-the-art read compressors and understand their shortcomings with respect to read compression. Finally, guided by the analysis, we present a simple algorithm for read compression geared towards achieving compression ratios close to the fundamental limit. The algorithm achieves compression ratios which are 1.3-2x better than the state-of-the-art compressors on the analyzed datasets. The algorithm compresses only the read sequence, works with unaligned FASTQ files, and does not need a reference.

The 10,000 Immunomes Project: A Data Resource for Human Immunology

Kelly Zalocusky

University of California, San Francisco

Despite increasing appreciation of the promise of clinical immunology in many areas of medicine, including cancer therapy, metabolism, and neurobiology, to date there is no large, representative data resource for the multitude of assays in human immunology. The ability to translate observations from model organisms to humans, the generation of new basic science hypotheses, and the interpretation of seldom-measured analytes are currently hampered by the lack of a reference human immunome. The 10,000 Immunomes Project aims to generate such a reference by compiling and harmonizing the measurements available on nearly 10,000 control subjects from the Immunology and Data Analysis Portal (ImmPort, www.immport.org), an archival repository for immunological research and clinical trials funded by NIAID. Data in the resource include ELISA, Luminex, flow cytometry, CyTOF, gene expression, HAI titers, clinical lab tests, and others. These data will be available through a web interface, allowing researchers to view, analyze, and download data across the 84 studies that contribute to the dataset. Our presentation will describe the demographics of the 10,000 Immunomes cohort, our data curation and standardization process, and the datasets we have successfully harmonized, including gene expression, ELISA, Luminex, and clinical lab tests. We will also present preliminary meta-analyses of these analytes. For example, we find that many soluble protein measurements vary significantly with age, gender, or ethnicity, implying that taking these variables into consideration will be important for the practice of clinical immunology and the advancement of precision medicine.

Whole genome sequencing of diverse human populations resolves causal regulatory variants

Michael Gloudemans

Stanford University

Expression quantitative trait loci (eQTL) studies have shed light on regulatory control of gene expression and provided an insight into the pathogenic mechanisms of human disease. However, since each regulatory locus contains many variants in linkage disequilibrium (LD), it is challenging to precisely fine-map the single causal variant responsible for an observed change in gene expression. To address this problem, we fine-mapped eQTLs using diverse haplotypes from six populations included in the 1000 Genomes Project. We hypothesized that variants fine-mapped using diverse populations would be more likely to represent true causal variants, and therefore would overlap with characterized and predicted functional regions. To test this hypothesis, we compared functional annotations of cis-eQTL sites before and after fine-mapping. We found that after fine-mapping, eQTLs were enriched for transcription factor binding sites and showed an increase in Eigen score, a predictive indicator of the pathogenicity of genomic variants. Furthermore, several loci that showed >99% LD in the UK10K Project, a collection of >1900 British individuals genomes, had these ties broken through fine-mapping, enabling greater precision for identifying causal variants. These results suggest that sequencing and analyzing genomes from a diverse set of ethnic backgrounds will provide additional power to detect functional genome variants such as eQTLs.

Automated methods for detection of axon bundle activation in Epiretinal Prostheses

Nandita Bhaskhar, Pulkit Tandon
Stanford University

Retinal prosthesis is an example of advanced electro-neural interfaces for treating blindness due to photoreceptor degeneration, which affects tens of millions of people worldwide. Epiretinal prostheses can produce more reliable visual percepts in blind patients via electrical stimulation of retinal ganglion cells (RGCs). Current epiretinal prostheses are prone to stimulation of axonal bundles, which relay the responses from 30-40 RGCs of different types to the brain. Indiscriminate stimulation of these bundles produces a signal that does not adhere to the natural neural code of the retina, and is known to significantly degrade quality of artificial vision in patients.

Our research aims to develop techniques in isolated retina for the design of advanced retinal prosthesis that delivers electrical stimuli more closely matched to the natural encoding of the retina. We use high-density multi-electrode arrays to electrically stimulate and record electrophysiological signals from primate retina.

We will discuss how we use these recordings to develop automated algorithms to determine whether an axon bundle has been activated, how it is validated and how it can be generalised to other Brain-Machine-Interfaces.

Tagging Patient Notes With ICD-9 Codes

Oliver Bear Don't Walk IV, Sandeep Ayyar, Manuel Rivas
Stanford University

There is substantial growth in the amount of medical/data being generated in hospitals. With over 96% adoption rate, Electronic Medical/Health Records are used to store most of this medical data. If harnessed correctly, this medium provides a very convenient platform for secondary data analysis of these records to improve medical and patient care. One crucial feature of the information stored in these systems are ICD9-diagnosis codes, which are used for billing purposes and integration to other databases. These codes are assigned to medical text and require expert annotators with experience and training. In this paper we formulate this problem as a multi-label classification problem and propose a deep learning framework to classify the ICD-9 codes a patient is assigned at the end of a visit. We demonstrate that a simple LSTM model with a single layer of non-linearity can learn to classify patient notes with their corresponding ICD-9 labels moderately well.

The Role of Alternative Splicing Regulation in the Innate Immune Response

Pratibha Jagannatha
University of California, Santa Cruz

The innate immune system is our first line of defense against infection. Initiation of the innate immune response involves a coordinated system of signaling pathways which results in an inflammatory response. While inflammation is important, chronic inflammation can lead to a variety of diseases. A number of these diseases, such as rheumatoid arthritis and cancer, are highly prevalent and currently incurable. Further study is required to identify better targets for more effective treatments. The goal of our study is to take a systematic high-throughput approach to rapidly identify genes that are critical for controlling inflammation. A recent study showed widespread RNA processing changes in macrophages, which are specialized cells in the immune system involved in identification and destruction of pathogens, upon infection. Alternative splicing occurs in over ninety percent of human genes and is an RNA processing mechanism that allows genes to code for multiple proteins through inclusion or exclusion of exons in mRNA. Our study focuses on

the role of alternative splicing regulation through the activation of specific pathways upon induced inflammation in macrophage cells. It involves stimulating macrophage cells with molecules that are recognized by toll-like receptors (TLRs), which are located in the membranes of macrophage cells. Specifically, we use lipopolysaccharide (LPS), Polyinosinic:polycytidylic acid (Poly I:C), Pam3CSK4, and R848, to investigate RNA splicing changes caused by specific TLR activation and associated pathways. We analyze RNA-seq data using DESeq2, to identify differentially expressed genes, and we use JuncBASE to identify differential splicing changes. Based on this data, we then identify candidate genes that show significant changes in gene expression and alternative splicing after induced inflammation. The majority of the significant events were categorized into the alternative first exon event type where the difference between the two isoforms is a shift in the first exon spliced into the mRNA. Using RT-PCR, we were able to validate three out of four selected genes: AMPD3, TNIP1, and RCAN1. These genes show alternative first exon event changes across most or all conditions. We are currently investigating whether these alternative first exon changes affect translation efficiency using polyribosomal profiling. Through this process, we will learn more about the different layers of gene regulation in the context of chronic inflammation.

Author Index