BCATS 2016

Compute Life

Abstract Book

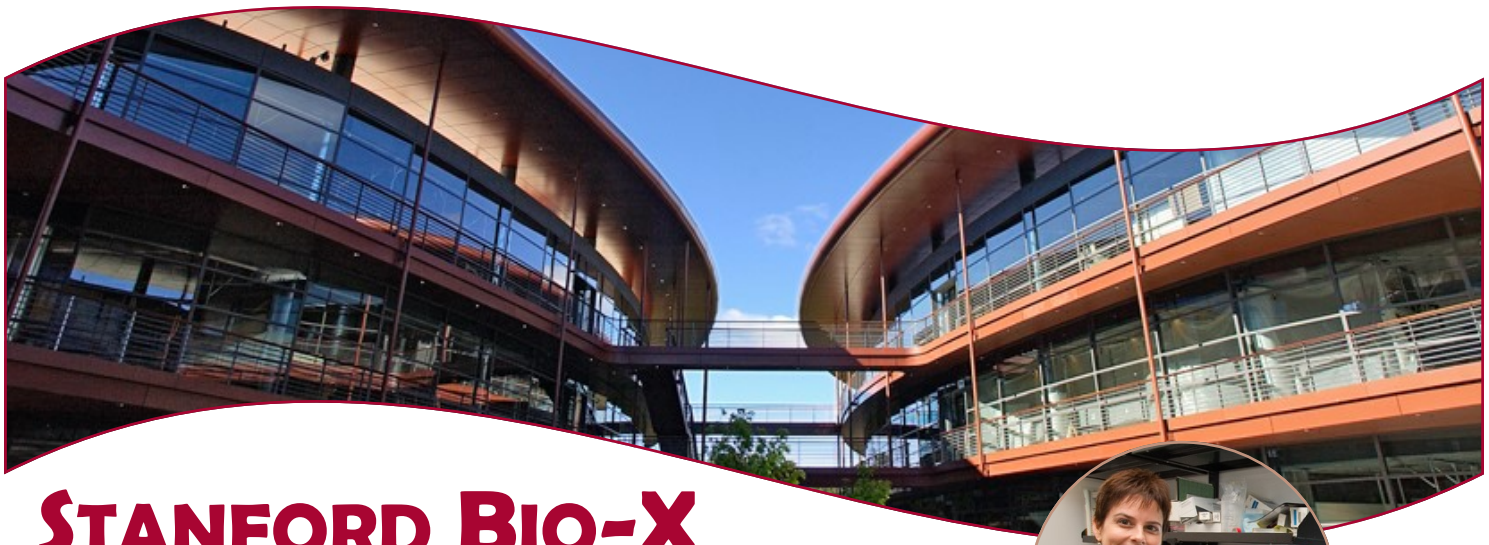
April 4th, 2016

16th Annual Biomedical Computation at Stanford Symposium (BCATS)

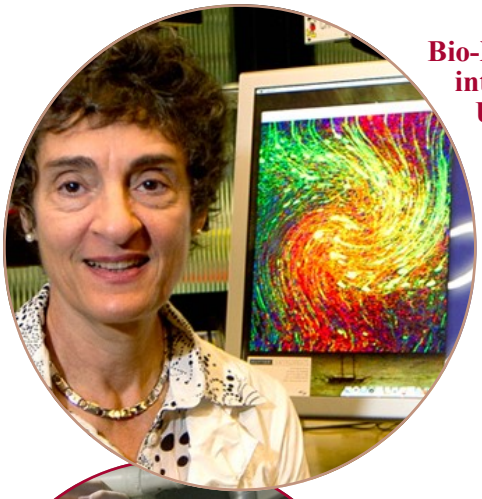
As a symposium that stands at the intersection of computer science, biology, statistics and medicine, BCATS has enjoyed continued growth over the last decade as the fields of biomedical informatics and computational biology gained traction. This year, researchers around the Bay area will attend BCATS to present novel applications in diverse fields such as genomics, proteomics, imaging and clinical informatics.

We would like to thank our keynote speakers, presenters, sponsors, and all 2016 attendees.

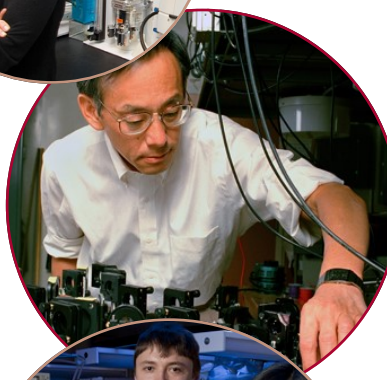
BCATS 2016 Sponsors



STANFORD BIO-X

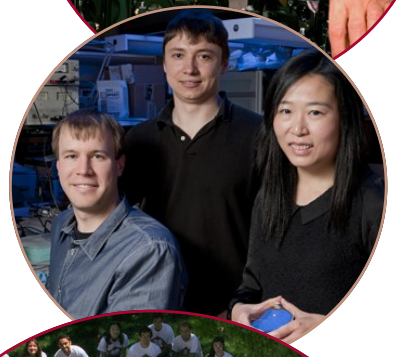


Bio-X is Stanford's pioneering initiative in the interdisciplinary movement at Stanford University. Over the past 16 years, Bio-X has developed into an **impactful institute that brings together biomedical and life science** researchers, clinicians, engineers, physicists, and computational scientists to unlock the secrets of the human body.



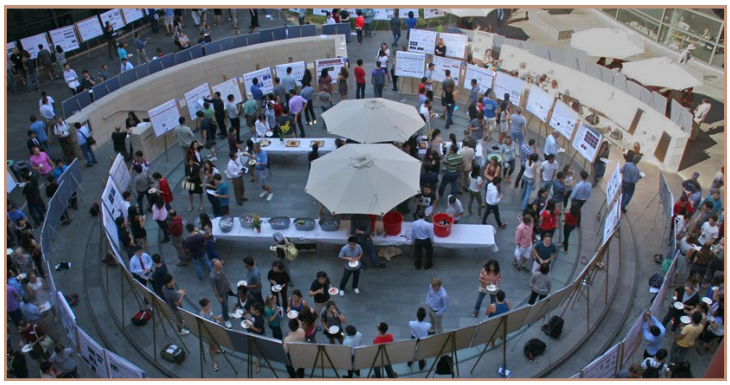
Over 700 Stanford faculty from all 7 Stanford schools and 60+ departments are affiliated with Bio-X, and participating in the following **Core Programs**:

- **Interdisciplinary Initiatives Seed Grant Program (IIP)**
- **Ventures**
- **Graduate Student Fellowships and Bio-X Stanford Interdisciplinary Graduate Fellowships**
- **Undergraduate Research Awards**
- **Travel Awards**



In addition, Bio-X establishes numerous **collaborations between the Stanford scientists and large corporations** by being a networking portal and providing opportunities for companies to join and participate in **symposia/seminars/mixers and customized technical summits, developing faculty liaisons**, and more.

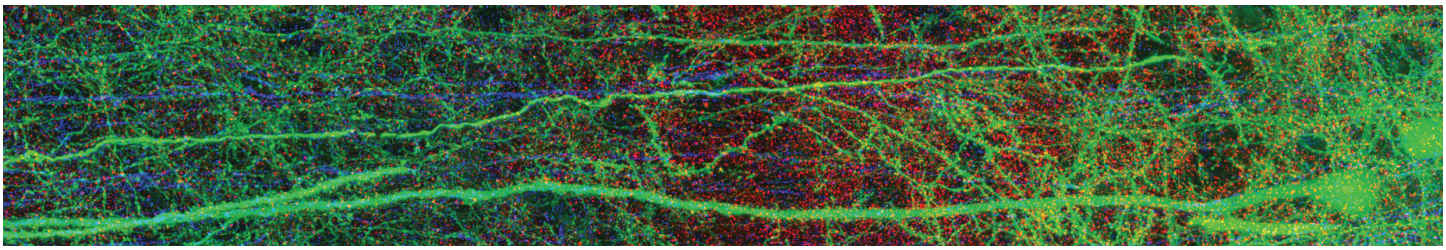
The **James H. Clark Center is the hub of Bio-X**, comprising the equipment, resources, and utilities required to conduct breakthrough research at the cutting edge of engineering, science, and medicine.



To learn more about Bio-X, visit our website:

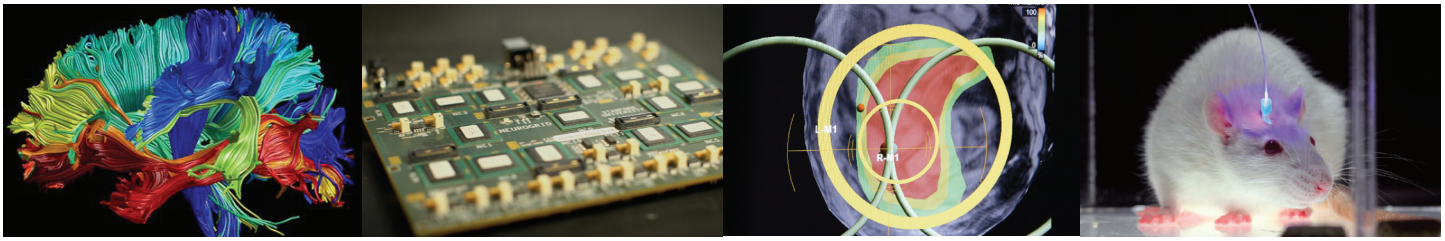
biox.stanford.edu

STANFORD BIO-X



Stanford Neurosciences Institute

We aim to understand how the brain gives rise to mental life and behavior, both in health and in disease. Our research community draws from multiple disciplines, including neuroscience, medicine, engineering, psychology, education and law. New discoveries will transform our understanding of the human brain, provide novel treatments for brain disorders, and promote brain health throughout the lifespan.



Big Ideas in Neuroscience

Interdisiciplinary teams of Stanford faculty dreamed big and imagined neuroscience transformed. The results were Big Ideas - new interdisciplinary collaborations spanning Stanford schools and departments aimed at tackling fundamental problems in neuroscience.

NeuroDiscovery - probing the inner workings of the brain

NeuroChoice - Probes how the brain makes decisions and expands that to influence public policy and economic decisions.

NeuroCircuit - Combines a detailed understanding of brain circuits with technology that modulates neural activity to develop improved ways of treating mental health conditions.

NeuroVision - Develops optical technologies that enable neuroscientists to visualize the brain in unprecedented detail.

NeuroEngineering - creating new technologies for interfacing with the brain

Brain-Machine Interface - Develops technology to interface with the brain and create intelligent prosthetics.

NeuroFab - Creates an incubator for next-generation neural interface platforms.

NeuroHealth - translating neuroscience discoveries into treatments

Brain Rejuvenation - Creates a center for neurodegeneration research focusing on brain maintenance and regeneration, and the role of the immune system in these processes.

Stroke Collaborative Action Network - Breaches barriers in our understanding of stroke to develop therapies and improve stroke recovery.

Opportunities

Interdisciplinary Research

Big Ideas in Neuroscience
Seed Grants
Stanford Program in Neuroscience and Society

Training

Interdisciplinary Scholar Awards
Stanford Interdisciplinary Graduate Fellowships
Software Carpentry Workshops

Engaging our Community

Weekly Seminar Series
Annual Symposium
Brain Broadcast Newsletter

Get Involved

http://neuroscience.stanford.edu

Stanford | Neurosciences Institute

Stanford Biomedical Informatics Training Program

The Biomedical Informatics Training Program (BMI) is an interdisciplinary graduate and postdoctoral training program, part of the Biosciences Program at Stanford University's School of Medicine. We offer MS and PhD degrees, and other coursework and research options.

Full-Time Graduate Programs:

- PhD in Biomedical Informatics
- Research MS degree (primarily, but not exclusively, for those with PhD and/or MD)

Part-Time Distance Education Programs:

- Professional/Honors Cooperative Program MS degree
- Certificate in Clinical Research Informatics or Bioinformatics
- Non-degree option for individual classes

For Stanford Students:

- Coterminal MS degree in Biomedical Informatics in addition to BA/BS
- PhD Minor in Biomedical Informatics for Stanford graduate students

For Stanford Medical Students:

- Coursework and research opportunities through Scholarly Concentration and Medical Scholars programs

All degree programs require rigorous coursework in Biomedical Informatics core courses, a coherent set of electives from Computer Science, Statistics, Math and/or and Engineering, and training in Social, Legal and Ethical issues.

For more information, contact:

Student Services Officer
Stanford Biomedical Informatics Training Program
Medical School Office Building, Room X-215
1265 Welch Road, Mail Code: 5479
Stanford, CA 94305-5479

Phone: (650) 723-1398
Fax: (650) 725-7944
email: bmi-contact@lists.stanford.edu
http://bmi.stanford.edu

Supported by National Library of Medicine Training Grant LM007033

Stanford ENGINEERING | Stanford Computer Forum

A Corporate Affiliates Program

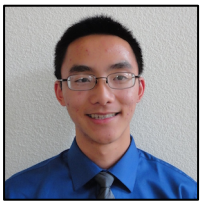
The Stanford Computer Forum is a cooperative venture that encourages collaboration between the Computer Science and the Electrical Engineering Departments at Stanford, and 100+ companies located in Silicon Valley, the rest of the U.S., Asia, and Europe. The Forum provides a mechanism for developing interaction with industrial researchers and their academic counterparts, promoting the exchange of the most advanced technological ideas in fields of computer science and electrical engineering. The Computer Forum also offers industry the opportunity to become familiar with the professional abilities and interests of Stanford students through its active recruiting program.

For further information please visit **http://www.forum.stanford.edu**
or e-mail forumstaff@cs.stanford.edu

BCATS 2016 Organizing Committee

Sandeep Ayyar
MS Student in Biomedical Informatics
ayyars@stanford.edu

Julie Chang
PhD Student in Bioengineering
jchang10@stanford.edu



Brian Do
Undergraduate Student in Computer Science
bdo@stanford.edu

Winn Haynes
PhD Candidate in Biomedical Informatics
haynewa@stanford.edu

Maulik Kamdar
PhD Student in Biomedical Informatics
maulikrk@stanford.edu

Stephanie Tzouanas Schmidt
PhD Student in Bioengineering
stephanie.schmidt@stanford.edu

BCATS 2016 Schedule

8:15 am	Badge Pickup and Breakfast
8:45 am	Opening Remarks
9:00 am	**Keynote Address:** Dr. Jonas Almeida (Page 11)
9:45 am	**Talk:** Scott Powers (Page 16) *Treatment effects for "patients like me"*
10:00 am	**Talk:** Avanti Shrikumar (Page 17) *Not Just a Black Box: Interpretable Deep Learning for Genomics and Epigenomics*
10:15 am	Coffee Break
10:30 am	**Keynote Address:** Dr. Laura Waller (Page 13)
11:15 am	**Talk:** Suzanne Tamang (Page 18) *Using Clinical Text Analysis to Enable Deep Phenotyping*
11:30 am	**Talk:** Natalie Telis (Page 19) *Functional characterization of Neanderthal introgression impacts on modern humans*
11:45 am	Lunch
1:00 pm	**Keynote Address:** Dr. Hao Li (Page 12)
1:45 pm	**Talk:** Amir Barati Farimani (Page 20) *Programming Graphene Nanopore with DNA Origami for DNA Sequencing*
2:00 pm	**Talk:** Can Cenik (Page 21) *Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans*
2:15 pm	**Poster Session** (Page 23)
3:45 pm	**Industry Panel:** Dr. Tiffany Chen, Dr. Ron Dror, Dr. Robert Gentleman, Dr. Geoffrey Rutledge (Page 14)
4:45 pm	Closing Remarks.

Keynote Talks

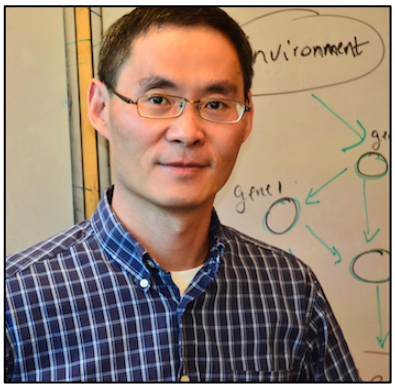
Dr. Jonas Almeida

Stony Brook University, New York

Professor and Chief Technology Officer,
Department of Biomedical Informatics

Dr. Almeida's lab researches at the intersection of semantic web abstractions and distributed cloud-computing approaches to bioinformatics application development in the pervasive web platform. This research often involves the application of mathematical modeling and machine learning for medical genomics, with a focus on The Cancer Genome Atlas (TCGA). Dr. Almeida is currently a Professor and Chief Technology Officer at the Biomedical Informatics Department of Stony Brook University (State University of New York, Long Island). Previously, he served as the inaugural director of a new Division in Informatics in the Department of Pathology of the University of Alabama at Birmingham (UAB) (2010-2014), and as Professor of Bioinformatics in the Division of Applied Mathematics of the University of Texas MD Anderson Cancer Center (2005-2010).

Dr. Hao Li



University of California, San Francisco

Professor, Department of Biochemistry and Biophysics

Dr. Li's lab studies the molecular mechanisms of aging and genetic determinants of complex human traits, using system biology approaches that combine quantitative experiments with theoretical modeling and bioinformatics analysis. His lab has developed microfluidic devices to analyze aging at the molecular level in single cells and genetic systems for high throughput lifespan screening. His lab has also developed various computational algorithms to analyze gene regulatory networks, and more recently to map genes that influence complex human traits by combining genome-wide association studies with other large-scale molecular trait data. Dr. Li is a Professor in the Department of Biochemistry and Biophysics at UCSF and the director of the UCSF Hillblom Center for the Biology of Aging. He received his BS in Physics from Peking University and Ph.D. in Theoretical Physics from MIT. He is a Changjiang Lecture Professor at Peking University, NIH transformative R01 awardee and Packard Fellow.

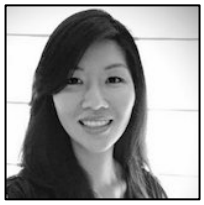
Dr. Laura Waller

University of California, Berkeley

Assistant Professor, Department of Electrical Engineering and Computer Sciences

Dr. Waller's lab develops new methods for designing imaging systems jointly in terms of hardware and software. This research is inherently interdisciplinary; drawing from expertise in optics, signal processing and computer science, with broad applications in bioimaging, defense, physical science and industrial inspection. Dr. Waller is an Assistant Professor at UC Berkeley in the Department of Electrical Engineering and Computer Sciences (EECS) and a Senior Fellow at the Berkeley Institute of Data Science (BIDS), with affiliations in Bioengineering and Applied Sciences and Technology. She was a Postdoctoral Researcher and Lecturer of Physics at Princeton University from 2010-2012 and received B.S., M.Eng., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT) in 2004, 2005, and 2010, respectively. She is a Moore Foundation Data-Driven Investigator, Bakar fellow, NSF CAREER awardee and Packard Fellow.

Industry Panel Speakers



Dr. Tiffany Chen

Director of Informatics,
Cytobank Inc.

Dr. Ron Dror

Associate Professor of Computer Science,
Stanford University

Dr. Robert Gentleman

Vice President of Computational Biology,
23andMe Inc.

Dr. Geoffrey Rutledge

Co-Founder and Chief Medical Officer,
HealthTap Inc.

Oral Presentations

Treatment Effects for "Patients Like Me"

Scott Powers

When considering potential treatment plans with their doctors, patients may well ask, "What is the expected effect of this treatment on patients like me?" More important to that individual than the overall mean treatment effect on the population is a prediction of the effect of the treatment in this particular instance. Our goal is to develop methodology to infer causal heterogeneous treatment effects on the basis of observational data. This is a challenging problem because for each patient we can observe only a response to treatment or a response to control, not both, and treatments are not randomly assigned as in an experiment. In this talk we present early-stage work on this problem, including our proposal of a novel decision tree, which makes splits on the basis of the statistical evidence of different treatment effects in the two resulting leaves. We extend the method by fitting random forests with these custom decision trees as the basis. We present simulation results to show in what settings this works, and in what settings it does not.

Not Just a Black Box: Interpretable Deep Learning for Genomics and Epigenomics

Avanti Shrikumar

Convolutional Neural Networks (CNNs) have emerged as a state-of-the-art technique for predicting transcription factor binding and epigenetic state at regulatory genomic regions from raw DNA sequence [1] [2] [3]. However, methods for interpreting these models leave much room for improvement. Weights of individual filters learned by the CNNs are often visualized and have been found to have some qualitative similarity to position weight matrix representations of regulatory sequence 'motifs' [1]. However, this approach fails to account for the inherently distributed representations learned by CNNs such that multiple filters often cooperatively represent a specific regulatory sequence element. Further, in-silico mutagenesis [2] or filter nullification [3] are commonly used methods to infer the relevance of filters or individual bases in input sequences. These approaches are very computationally expensive. Moreover, we show that they can often provide misleading results when the input sequences contains redundant signals that potentially buffer each other. Here, we present DeepLIFT (Deep Linear Importance Feature Tracker), a family of novel techniques that circumvents the aforementioned limitations, and, to our knowledge, is the only technique that can integrate the combined effects of multiple cooperating filters. We demonstrate the successful application of DeepLIFT to interpreting deep learning models for diverse learning tasks in regulatory genomics and epigenomics, including the prediction of chromatin states and histone marks from ATAC-seq, DNase-seq and MNase-seq data as well as prediction of in-vivo binding sites of transcription factors by learning from TF ChIP-seq data. We show that DeepLIFT not only produces much cleaner and broader sequence motifs than traditional approaches, but also enables deconvolving heterogeneity of regulatory sequence grammars. Our framework provides the first comprehensive interpretation engine for deep neural networks in genomics and epigenomics.

[1] DeepBind (Nat. Biotechnol. 33,831–838, 2015)
[2] DeepSEA (Nat. Methods 12, 931–934, 2015)
[3] Basset (bioRxiv, 10.1101/028399, 2015)

Using Clinical Text Analysis to Enable Deep Phenotyping

Suzanne Tamang

Metastatic breast cancer takes the lives of 40,000 individuals in the US annually and the patient, family and population health burden are large. Although the Surveillance, Epidemiology and End Results (SEER) registry has enabled numerous population-based studies of breast cancer patients, SEER does not capture distant recurrence events. Moreover, metastatic recurrence is not documented as a structured field in medical records or billing codes.

Detailed phenotyping of an individual's observable traits, increasingly referred to as "deep" phenotyping, seeks to improve the level of detail captured in electronic phenotyping. Using the Stanford and PAMF's Oncoshare database, a multi-institutional EHR and SEER linked database of breast cancer patients in California, we sought to improve on the capture of current breast cancer phenotypes to allow investigation of distant recurrence events.

Our innovation is an informatics approach that includes natural language processing (NLP) to identify metastatic breast cancer patients – without the need for an expert-annotated development corpus. Evaluation of our system's performance on detecting distant recurrence events documented in patient notes showed an F1 value of .93 for our rule-based extractor and .91 for our statistical extractor; patient-level evaluation indicated 76% sensitivity and 96% specificity for the rule-based extractor and 92% and 83% sensitivity for our statistical extractor. Our future work seeks to address clinically relevant questions about the survival outcomes for HER2 positive and TNBC subtypes, and to characterize the changing face of metastatic breast cancer in a real-world treatment context across two healthcare systems.

This talk will describe the process we used to develop a custom tool to extract distant recurrence events documented in patient notes using CLEVER (for Clinical EVEnt Recognizer), which is available with a BSD license. Our work provides additional evidence for the ability of clinical text analysis tools to improve on electronic phenotype definitions. Our findings suggest that when a small set of targeted information tasks can be identified, modern information extraction methods for building statistical extractors using clinical text may have the potential to replace traditional clinical NLP paradigms that are costly, labor intensive, and otherwise resource-prohibitive.

Functional Characterization of Neanderthal Introgression Impacts on Modern Humans

Natalie Telis

There has been a long history of interest in the extent of, and selective consequences of, human interbreeding with other hominins. Most efforts in human paleogenomics have focused on characterizing the population structure of archaic hominins, as well as the pattern of interbreeding with modern humans. As a result of the fraction of introgression observed in modern populations, it is of great interest to characterize functional and clinical consequences of introgression. Though recent work has characterized select examples of both adaptive and pathogenic variation, it remains of great interest to characterize the broader, functional nature of Neanderthal introgression and its role in modern human phenotypes. We create a pipeline to characterize functional patterns of introgression, and infer relationships between these patterns and the selective pressures on humans using a novel metric of selection. We find evidence for tissue-specific patterns of introgression genome-wide, with potential impacts on broader phenotype. We conclude that the array of selective pressures on Neanderthal variants is diverse, and that functional annotation constitutes a novel path to understanding the interaction of genetic variation and complex traits.

Programming Graphene Nanopore with DNA Origami for DNA Sequencing

Amir Barati Farimani

DNA origami nanostructures can be used to functionalize solid-state nanopores for single molecule studies. In this study, we characterized a nanopore in a DNA origami-graphene heterostructure for DNA detection. The DNA origami nanopore is functionalized with a specific nucleotide type at the edge of the pore. Using extensive molecular dynamics (MD) simulations, we computed and analyzed the ionic conductivity of nanopores in heterostructures carpeted with 1 or 2 layers of DNA origami on graphene. We demonstrate that the nanopore in DNA origami-graphene gives rise to distinguishable dwell times for the four DNA base types, whereas for the nanopore in bare graphene, the dwell time is almost the same for all types of bases. The specific interactions (hydrogen bonds) between DNA origami and the translocating DNA strand yield different residence times and ionic currents. We also conclude that the speed of DNA translocation decreases due to the friction between the dangling bases at the pore mouth and the sequencing DNA strands.

Integrative Analysis of RNA, Translation, and Protein Levels Reveals Distinct Regulatory Variation across Humans

Can Cenik

Elucidating the consequences of genetic differences between humans is essential for understanding phenotypic diversity and personalized medicine. Although variation in RNA levels, transcription factor binding, and chromatin have been explored, little is known about global variation in translation and its genetic determinants. We used ribosome profiling, RNA sequencing, and mass spectrometry to perform an integrated analysis in lymphoblastoid cell lines from a diverse group of individuals. We find significant differences in RNA, translation, and protein levels suggesting diverse mechanisms of personalized gene expression control. Combined analysis of RNA expression and ribosome occupancy improves the identification of individual protein level differences. Finally, we identify genetic differences that specifically modulate ribosome occupancy—many of these differences lie close to start codons and upstream ORFs. Our results reveal a new level of gene expression variation among humans and indicate that genetic variants can cause changes in protein levels through effects on translation.

Poster Presentations

BCATS 2016 Accepted Posters

Pancancer Analysis of DNA Methylation Using Methylmix

Marcos Prunello

DNA methylation holds important epigenetic regulatory effects and hypo and hyper-methylation are known to play an important role in cancer. We applied MethylMix, a method that identifies differentially methylated and transcriptionally predictive genes, in a pancancer study of 26 combined cancer sites, with about 10000 samples from The Cancer Genome Atlas. We found 146 genes with differential methylation in cancer samples with respect to normal. We characterized each cancer sample according to its methylation profile for these genes and then clustered the patients across all cancers. Consensus clustering identified 15 pancancer clusters. A subset of these clusters are mainly constituted by samples from one cancer site or by samples from related cancer types, but others consisted of samples from different tissues reflecting commonalities in their methylation profiles. Clusters associated to brain tumor showed the highest number of hyper-methylated genes and clusters related to ovarian and thyroid cancer, the lowest. Survival was significantly different among clusters but also between samples of the same tissue that were grouped into different clusters. We analyzed clinical features and found significant differences in terms of age, smoking, gender, stage, grade, pathological spread of the disease and histology. For some cancers the partition of samples of the same tissue in different clusters was associated to gene expression based subtyping. We used SAM to identify over-expressed genes in each cluster, further investigated with GSEA. Several sets of genes known for being regulated by methylation showed enrichment and in most cases the set of up-regulated genes in a particular cluster was enriched with sets associated to the predominant cancer site in that cluster. This analysis reveals new similarities of malignantly transformed tissues based on common methylation patterns and might be useful to redefine cancer treatment based on methylation pattern instead of anatomical site of origin.

A Chronic Lymphocytic Leukemia Classification System Based on Normal B-cell Subset Associated Gene Signatures (BAGS)

Caroline Hol Noergaard

CLL is a heterogeneous B cell malignancy characterized by the accumulation of mature, neoplastic B-lymphocytes in the peripheral blood, bone marrow, and secondary lymphoid organs. Patients with CLL undergo highly variable disease courses, ranging from chemotherapy-resistant and fatal disease shortly after diagnosis to slow progression and survival for decades. Today's diagnostic tests for chronic lymphatic leukemia (CLL) reflect the criteria of the updated WHO classification based on biomarkers and clinicopathologic heterogeneity. To that end, we propose a new classification of malignant B cell diseases, by a B-cell subset associated gene signatures (BAGS) classifier generated from the normal B-cell hierarchy in lymphoid organs. Our hypothesis is that BAGS subtyping for CLL can provide prognostic value and new biologic insight.

We combined fluorescence-activated cell sorting (FACS) and gene expression profiling (GEP) using the Human Exon 1.0 ST array platform to generate BAGS classification for the following B cell subsets in healthy bone marrow: Pre-BI, Pre-BII, Immature, Naïve, Memory, and Plasma Cell.

Construction of the classifier was based on median-centered probe sets from the bone marrow data using cross-validated regularized multinomial regression with six discrete outcomes representing BAGS. We identified nine publicly available cohorts consisting of Affymetrix U133 plus 2.0 microarray data gathered across geographical regions and time periods from more than 1500 untreated CLL patients. Each patient underwent BAGS assignment according to the highest predicted probability score.

The resultant BAGS assignments to the CLL samples exhibited similar BAGS subset frequencies across all CLL cohorts. The BAGS groups were subsequently divided into a pre- and post-germinal subgroup. Preliminary analyses showed a significant ratio of IgVH unmutated samples in the pre- germinal subtype group when compared to the post-germinal subtype group. The prognostic impact of BAGS was analyzed using overall survival data from 107 CLL patients, and showed a tendency for later differentiation BAGS subgroups to have poorer prognosis compared to earlier differentiation BAGS subgroups, although no statistically significant difference could be observed. Furthermore, analyzing differentially expressed genes between the BAGS subgroups are currently being pursued. Comparing the BAGS subgroups will enable us to improve the understanding of the reported heterogeneity in the cell of origin of CLL.

A Computational Pipeline to Uncover Genomic Regulatory Sequences that Modulate the Wnt Signaling Pathway

Bhaven Patel

The Wnt signaling pathway is crucial to development and its components are frequently mutated in variety of cancers. Although the core components of the canonical Wnt signaling pathway have been discovered, many regulators of the pathway remain undiscovered. To identify novel genes involved in Wnt signaling, we performed unbiased forward genetic screens in a human haploid cell line using retroviral mutagenesis. Analysis of these screens has traditionally been “gene-centric”, focused on mapping the insertions in genomic regions with annotated, protein-coding transcripts. We hypothesized that recurrent insertions in non-protein coding regions of the genome may identify regulatory elements or cryptic transcripts that regulate Wnt signaling. Therefore, we developed a new bioinformatics pipeline designed to be “gene blind”. Our pipeline identifies recurrent retroviral insertions in 3 million one thousand base-pair bins in the human genome, chosen without regard to gene boundaries. Using this pipeline, we have found several genomic regions, including unannotated areas near the genes LRP6, TCF7L2, TFAP4 and APC, which could play important regulatory roles in Wnt responsiveness. We believe that such genomic regions represent regulatory elements, such as enhancers and cis-acting non-coding RNAs. In summary, our pipeline promises to probe the genomic regulome of a signaling pathway that plays a central role in development, cancer and regeneration.

Methods for Differential Abundance Estimation on Sparse Microbiome Data

Lan Huong Nguyen

The field of microbial ecology has undergone a major transformation with the advent of highly parallel next generation sequencing (NGS) technologies. The advances have facilitated the discovery of many microbial species. These improvements are of course accompanied by many challenges. DNA sequencing data now consists of discrete, highly skewed counts of sequence reads. Thus, the matured statistical methods designed for continuous microarray data are no longer applicable. The objective of many microbial compositional studies is to determine species differentially abundant in samples subject to two or more conditions. It is a common practice for researchers to employ frameworks, many based on a negative binomial model, designed for differential expression of RNA-seq data to perform the differential abundance analysis on the microbiome data. While both microbial and RNA-seq data both consist of non-negative integer values, their distribution and sparsity levels are very different. Zero counts are extremely frequent in the microbial community surveys and rarely occur in RNA-seq data. We propose a new framework based on a Zero-Inflated Negative Binomial model to account for this difference. The method makes use of the existing normalization and regularization tools to correct for uneven sampling depths and help alleviate over-parametrization, which are issues shared by both microbial and RNA-seq data. We compare our method with the existing DE RNA-seq packages, edgeR and DESeq2 and another software developed specifically for microbiome data, metagenomeSeq, which is based on a Zero-Inflated-Gaussian model. The performance in terms of sensitivity and specificity is evaluated on simulations based on the Human Microbiome Project dataset subject to controlled levels of sparsity.

A Deep Learning Model for Predicting and Understanding CTCF Binding across Cell Types

Irene Kaplow

CTCF is involved in holding DNA loops together, thereby playing a key role in determining how genes are regulated. Thus, we can study the mechanism behind gene expression variation across cell types and individuals by understanding how CTCF binding varies. However, our current understanding of the relationship between DNA sequence and CTCF binding is incomplete. We have therefore developed a deep neural networks approach to predict CTCF binding across individuals and cell types. We show that we can predict both the presence of a CTCF peak and the height of that peak in multiple cell types. We are currently interpreting these models to understand the sequence signatures underlying binding in different cell types. We plan to use these comparisons to understand how DNA looping differs across cell types, which we anticipate will provide a mechanism to explain expression differences.

Predicting Rheumatoid Arthritis Patient Outcome Using Lab Tests

Alice Yu

Rheumatoid arthritis (RA) is an autoimmune disease that is currently affecting 1.3 million Americans and the number of cases is still on the rise. It leads to irreversible joint damage and needs to be treated immediately to delay disease progression. Moreover, RA is commonly left undiagnosed because early symptoms resemble other diseases and the lab tests used for diagnosis lack conclusive results. Instead, the current procedure involves a variety of factors that are combined into a diagnostic patient score and only detects RA at more severe stages. To reduce the uncertainty of this method, RA lab tests have been examined to determine which ones are most predictive of RA patient outcomes. Identifying predictive lab test(s) will provide guidelines for clinicians as to which tests are relevant to order and how the tests should be interpreted.

To find which tests are most deterministic of varying RA stages, we leveraged RA patient data from the STRIDE database and determined patient outcome using Erythrocyte Sedimentation Rate (ESR) and C-Reactive Protein (CRP) tests, which are used as disease progression metrics in current clinical settings. We created classifiers to predict patient outcomes and determined that Mean Corpuscular Hemoglobin Concentration (MCHC), Red Cell Distribution Width (RDW), and Hemoglobin counts (HGB) proved to have the most predictive values. All these measurements are found in the complete blood count (CBC) tests, whereas the additional lab tests, which patients took, proved to be the least predictive. Knowing which tests have a higher chance of predicting RA outcome increases early RA detection and yields further insight on the nature of the disease and its underlying physiological mechanisms.

Diagnosis of Heart Diseases via CNNs

Yanyang Kong

We applied Convolutional Neural Networks to predict end-systolic and end-diastolic volumes of left ventricle, which can be used to compute ejection fraction, an important indicator of heart disease. We used the MRIs provided by National Institutes of Health and Children's National Medical Center as our data set, and built an axis-based model that is robust to MRIs in different axis planes and in different qualities.

Automated Bone Age Classification with Deep Neural Networks

Matthew Chen

In this project we look at the use of Convolutional Neural Network methods to train a model to predict developmental bone age of a patient given x-ray images. Bone age assessment is a standardized process by which a medical practitioner determines the skeletal maturity of a child through a scan of their hand. It is commonly used in comparison with chronological age as an indicator for developmental issues for a child. It is also useful in determining age where birth records are not accessible. The standard test for bone age assessment involves a radiological scan of the left hand, which is then manually compared to an atlas of reference images. Automated methods for bone age assessment have been proposed in the past. These methods generally involve segmenting the scan into regions of interest and running a classifier on the results. In this paper we aim for a more general approach where we avoid creating hand crafted features by training a convolutional neural network directly on the input pixels.

We use the Digital Hand Atlas dataset that is composed of scans independently annotated by two radiologists. Previous methods for this task generally involve a pipeline of segmentation and hand crafted feature extraction. We look to move away from this approach given recent advances in the effectiveness of convolutional neural networks for image classification.

We find that using a convolutional neural network approach for this image classification task, we are able to achieve a top one and two accuracy of 46% and 70% respectively with root mean squared error of 1.1 years, on our validation set. We observe that our largest jump in accuracy resulted from augmenting our dataset with random distortions. This seems to indicate that the performance is largely dependent on the number of training examples and would likely see further improvement with more data. Our work adds to the growing evidence that training deep neural networks can be effective for many nuanced medical image classification tasks that were once in the domain of manual feature engineering.

Artificial Control of Soft Surgical Robots with Gaussian Processes

David Eng

The use of robotic devices for surgical interventions has gained significant traction in medical facilities in the last 25 years. Conventional interventions, such as the da Vinci Surgical System, use rigid materials to fabricate these robotic systems, which are easily represented as rigid members connected at discrete joints.

However, rigid robots like the da Vinci system present severe limitations. Though their rigid frames allow for precise and predictable control of the system, these frames also introduce significant risks in surgeries involving highly sensitive and vital organs, such as the lung or heart.

Since they bend and twist with high curvatures, soft robots offer the desired properties of a more versatile tool. In surgeries such as cardiac ablation for atrial fibrillation or diagnostic biopsies of the lung, the use of soft robots instead of rigid ones dramatically reduces the risk of inadvertent damage to impacted organs.

Despite their exciting surgical implications, soft robots lack widespread adoption due to a lack of accurate control schemes capable of governing their movements. Existing models struggle with the high dimensionality of the state space through which these robots must traverse, as well as the amount of noise introduced by the randomness of the surrounding organs.

To improve the viability of soft robotics in medical facilities, we propose an online reinforcement-learning algorithm called Gaussian Process TD (GPTD) Learning to control a soft surgical robot catheter in the noisy environment of the human body. In particular, we define a probabilistic generative model to estimate the value function in continuous spaces and introduce an online sparsification to allow for a computationally tractable update.

The online GPTD algorithm can achieve a policy within 10 percent of optimal with an update cycle time of less than 14 ms. Overall, we conclude that the GP approach performs well for single-point trajectories and remains a promising way to control soft robots in actual surgical interventions.

Using Music to Analyze Protein Sequences

Aaron Kosmatin

Proteins are typically represented as strings of characters encoding amino acids. Humans process knowledge and information in a variety of different ways; one of them being auditory. There are similarities between the structures of proteins and music; both are composed of phrases organized into themes [1].

The first aim is to convert amino acid sequences to music. We want to generate pleasant music while preserving relationships between amino acids properties. A good musical mapping makes it easier to listen to what is generated, and easier to remember repeated sections. Unlike previous mappings that have missed musical structures, we introduce instrumentation commonly found in pop music. Our keyboard mapping of amino acids to musical notes is embellished by using drums to dictate the size of the amino acid, by guitar to differentiate between charged, uncharged, aliphatic, and aromatic amino acids and lastly by four different melodies to differentiate between non-polar, polar, positively and negatively charged amino acids. Our mapping follows common musical structures, making biology, music, and computer science more appealing to a wider range of audiences.

The second aim is to analyze protein sequences that have been converted to music according to the mapping described above. We create a database of recordings of orthologs: similar proteins found in different organisms. The database will also contain recordings of pairwise alignments of orthologous proteins. By juxtaposing and recording two orthologous proteins simultaneously, but with different instrumentations, one can hear the conserved regions between the two proteins. The conserved region is a theme and the various orthologous proteins that contain this theme, play it as variations on that theme.

Lastly, we plan on extending this work to the study of protein pattern recognition. Rather than running hidden Markov Models (HMMs) built for the analysis of protein sequences [2], we plan to run original HMMs from speech recognition on the musical encodings of protein sequences and then perform inverse mappings to identify patterns in proteins. We believe that this approach will yield better results since HMMs were originally designed for speech recognition.

[1] Dunn J. Clark M. (1999) Life Music: Sonification of Proteins. Leonardo V.32(1):25-32.
[2] Durbin R. et al., (1999) Biological Sequence Analysis: Probabilistic Models of Protein & Nucleic Acids. Cambridge University Press.

Cardiovascular Edge Detection using Neural Networks

Gabriel Maher

In order to perform computational fluid dynamics simulations of blood flows in patients, models of their cardiovascular system are required. Currently the only way to obtain models of sufficient quality is for practitioners to manually construct them from volumetric scan data using computer aided design (CAD) software. In order to automate the process of cardiovascular model construction it is necessary to be able to produce high accuracy edge-maps of the cardiovascular system from volumetric scans.

Despite numerous advances in the area of deep learning applied to edge-detection and image segmentation, current techniques have not yet seen application towards cardiovascular edge-detection. In this paper we make use of convolutional neural networks to develop a classifier that can segment volumetric scan data to detect blood vessels and other cardiovascular structures in volumetric scan data. Specifically a classifier is developed that can identify which pixels of a 2D image are blood vessels. Randomized search techniques are used to determine an optimal neural network architecture and training hyperparameters. The final network architecture is able to achieve 89% accuracy on a test set of 5,000 images.

Systematic Comparison of Genome-Wide CRISPR/Cas9 and Rnai Screens for the Identification of Essential Genes

David Morgens

Genome-wide perturbation screens in mammalian cells using RNAi and more recently CRISPR/Cas9 systems have been extraordinarily powerful, yet exhibit high variability among reagents, which complicates interpretation of screen results. We have developed a novel statistical framework, Cas9 High Throughput Maximum Likelihood Estimator (casTLE) to provide both a confidence score as well as an estimate of true effect size that accounts for this heterogeneity. Here we have compared the ability of shRNA and CRISPR/Cas9 screens to identify essential genes in the human chronic myelogenous leukemia cell line K562. Using a gold standard set of essential genes and casTLE, we found that the precision of the two libraries was comparable and that combining data from both screens noticeably increased performance. Strikingly, results from the two screens show little correlation, which can be partially explained by identification of distinct essential biological processes with each technology.

Transport of Subcellular Cargo faces a Critical Speed-Precision Tradeoff in Neurons

Alex Williams

Neurons are morphologically unusual cells: they are large, have elaborately branching processes and show spatially polarized expression of proteins, mRNAs and other subcellular components. Since all gene expression must begin with transcription in the nucleus, neurons must rely on a complex active transport network to deliver all cargo to appropriate locations throughout the cell. How neurons solve this considerable logistical problem remains poorly understood.

Subcellular cargo shows stochastic bidirectional movements when transported along microtubules by kinesin and dyenin motor proteins. Signals such as synaptic input are known to locally affect the trafficking and recruitment of subcellular cargo, suggesting that cargo searches a neuron's morphology by a noisy, random walk before being captured in a subcellular region. Preeminent models of long-lasting synaptic plasticity — an important molecular mechanism that supports learning and memory, have invoked this intuitively plausible model. However, there have been few attempts to determine whether local signals can orchestrate a given spatial distribution of cargo in realistic dendritic morphologies.

We mathematically formalize this conceptual model and present three central findings: 1) cargo can indeed be distributed in proportion to local demand in arbitrarily complex morphologies, given sufficient time. 2) The same distribution of cargo can be achieved by qualitatively different transport strategies, which provides insight into seemingly contradictory experimental observations. 3) There is an critical and unavoidable tradeoff between the speed and precision of cargo delivery. The model predicts delays of many hours or even days for modestly accurate cargo distribution — substantially slower than the requirements of certain plasticity models. Intuitively, this tradeoff arises because facilitating local capture diminishes exploratory movement of cargo across the dendritic tree.

Given the likely importance of this speed-precision tradeoff for neural function, we explored ways in which it might be circumvented. If the distribution of cargo is highly stereotyped, we show that both fast and precise transport can be achieved by introducing global rules such as spatial gradients in the trafficking kinetics of motor proteins. However, these tuned mechanisms are fragile to perturbations in the spatial demand for cargo and other sources of noise. Hence any attempt to beat the speed-precision tradeoff comes at the cost of inflexibility.

Together, these results suggest fundamental limits to soma-to-synapse trafficking that constrain plausible models of synaptic plasticity, homeostasis and neuronal function.

Cutting Angle Selection based on Histogram of Oriented Gradients and Spatial Constraints

Jing Xiong

A longstanding problem in neuroscience research has been the need to manually annotate brain regions of labeled neurons in order to recover the brain neuron circuitry. This requires pixel-wise alignment of a sequence of experimental slices to a reference brain volume. However the generation of high-resolution mouse brain microscopic images often introduces artifacts and noise that makes the 3D reconstruction of the experimental mouse brain difficult, and therefore 3D registration programs cannot be directly used. More often neuroscientists find a matching section in the annotated reference brain volume that matches an experimental slice and manually annotate the brain region of an annotated neuron or use a 2D registration program. Many existing 2D non-rigid registration methods have shown good results with/without human initiated grid control points. However manual cutting plane selection can be very inaccurate due to cutting angle difference, artifacts, and brain structural variation, and manually generated result can hardly be consistent. Our automatic program finds the best cutting angle difference between a sequence of experimental slices and the Allen Mouse Brain Reference Atlas with the Histogram of the Oriented Gradients and spatial constraint. Scientists can facilitate the detected best cutting angle to reslice the Allen Mouse Brain Atlas in the same sectioning angle as the experimental volume. Results on multiple brains generated by the program have been evaluated and consented by biologists.

Modeling ancestry-dependent phenotypic variance reduces bias and increases power in genetic association studies

Shaila Musharoff

Many complex human phenotypes vary dramatically in their distributions between populations. Genetic association studies typically use estimates of population structure due to shared ancestry, such as principal components (PCs), as fixed-effect covariates to prevent confounding caused by a dependence of phenotypic mean on ancestry. However, the standard approach of including PC covariates in linear regression models assumes that different populations have the same phenotypic variance, which may not hold for recently admixed populations or disease phenotypes. In this work we consider the possibility that populations with differences in phenotypic mean also have differences in phenotypic variance. First we show this is the typical case under an additive genetic architecture. Then we develop a new likelihood-based method, based on a double generalized linear model, to account for relationships between ancestry and phenotypic variance in genetic association studies. In simulations, our test has better power than several linear regression tests that assume equal variance across groups. We observe power increases of 12 - 66% and obtain unbiased parameter estimates for data simulated with realistic human disease effect sizes, population minor allele frequency (MAF) differences of 0.45, and a three-fold difference in variance. Furthermore, we show that the current gold standard approach of linear regression with PC covariates can lead to inflation or deflation of p-values for tests of genetic association. For example, simulated data where an allele is rare in one population (MAF = 0.01) and common with three times as much variance in another (MAF = 0.5), produces test statistics with an inflation factor (lambdaGC value) of 1.56, which is corrected to 1.02 (i.e. almost no inflation) by our method.

We apply our method to admixed individuals from the Study of African Americans, Asthma, Genes and Environments (SAGE) and find significant associations of baseline lung function (i.e. before treatment with asthma drugs) with estimated global ancestry when either sex or BMI (body mass index) is a sole covariate. In contrast, standard linear regression (which assumes equal variance among groups) requires sex, BMI, and additional covariates (age, height, and weight) for significance. Our method finds significant associations with fewer covariates because phenotypic variances as well as phenotypic means differ among groups. As this is characteristic of other phenotypes, or method is promising for association studies and other data with unequal group variances, such as gene expression data.

Role of Glycans in NMDA Receptors: Computational Approach

Anton Sinitsky

N-methyl-D-aspartate receptors (NMDARs) are transmembrane proteins expressed in nerve cells that play a key role in memory formation and learning. It is known that NMDARs in vivo are heavily glycosylated. However, how glycosylation state affects the structure and function of NMDARs is largely unknown. Previous computational studies have omitted glycans, putting into question the relevance of such models for understanding the behavior of NMDARs in vivo. We have performed molecular dynamics (MD) simulations of the GluN1 ligand binding domain (292 amino acid fragment) of an NMDAR in a physiologically relevant glycosylated state, namely with three covalently attached Man5 glycans. As a control, simulations of a glycan-free protein were performed. In total, MD trajectories with a total duration of 62 μs (plus 58 μs for the control system) were generated. Markov state models were used to interpret the results of simulations. We find that the main effect of glycosylation emerges from transient (on the timescale of ~10-100 ns) interactions between Man5 glycan at residue Asp48 (corresponding to residue 440 in the full GluN1 subunit sequence), on the one hand, and the protein fragment with residues 202 to 215 (or 710 to 723 in the full GluN1 sequence) on the other hand. These noncovalent interactions shift the ensemble distribution of the ligand-binding domain of NMDAR towards more closed conformations, that is, in the same direction as binding glycine (or another coagonist). For this reason, we predict that the glycan at GluN1-Asp440 plays a coagonist-like role, and that an untypical glycosylation state of this site may lead to schizophrenic symptoms. In summary, computational modeling of NMDARs provides a unique source of information about the structure of glycosylated receptors and their submicrosecond dynamics, hardly available from modern experimental approaches.

Quantifying Mammalian Learning: Large-Scale Detection of Dendritic Spines

John Lambert

We demonstrate that by training a Convolutional Neural Network (CNN) on a sliding window and by altering the distribution of classes we can predict the location of dendritic spines in microscopies of stained mouse neurons. As the main receivers for synaptic connections, dendritic spines are of major importance to neuroscience, reflecting cognitive ability and disorders. We show that, for our specific task of detecting the location of spines on a dendrite, “transfer learning” is not well-suited to the task, despite the success achieved by the YOLO and Faster-RCNN algorithms on the PASCAL VOC 2007, MS COCO, and ILSVRC Image Detection datasets. Our task is complicated by incomplete gold standard labelings; to make progress without editing the ground truth labeled by neuroscientists, we simplified our problem into a sliding window classification on down-sampled data. Our simple three-layer CNN shows strides towards replacing laborious hand-annotation of dendritic spine structures with automatic, algorithmic detection. Not only does our classifier demonstrate 100% recall on training data, it also succeeds in finding spines not originally labelled by humans. We demonstrate that when trained upon an input label of one correct spine location per image, our model can detect up to seven times that many true spine locations in that same image, at test time. Our goal is to build out a model that can be highly scalable and useful for researchers in the future.

A Framework for Automated Tuning and Uncertainty Quantification in Multiscale Coronary Flow Simulations

Justin Tran

Computational simulations of coronary flow can provide non-invasive information on hemodynamics that can aid in clinical treatment planning and disease progression research. In this study, patient-specific geometries are constructed and combined with finite element flow simulations using the open source software SimVascular. Lumped parameter networks (LPNs), consisting of circuit representations of hemodynamic behavior, can be used as coupled boundary conditions for the flow solver to model the global hemodynamic behavior of the vasculature not captured in the 3D model. The parameters of the LPN are tuned so the outputs match a patient's clinical data such as blood pressure, ejection fraction, etc. There are, however, hurdles in applying this framework to model large numbers of patients. Currently, setting up the necessary files to run simulations for a single patient requires a significant amount of user knowledge and manual interaction. After producing the geometric model, users previously had to tune the parameters of the LPN to match patient targets, a laborious and time consuming process. Additionally, results from these simulations are typically reported as deterministic without any associated uncertainty, even though the clinical, physiologic, and image data used to tune these parameters have uncertainty associated with their measurement. To address these concerns, we propose a framework to automate the parameter estimation process needed for automated tuning, and provide tools to perform forward uncertainty propagation. This framework can directly read in all the required information from the geometric model to build the system of equations governing the LPN for each individual patient, and produce the input files needed for simulation. To address the issue of parameter tuning, we combine Bayesian registration with Nelder-Mead optimization to produce optimal parameter values that match the patient's clinical data. This Bayesian framework utilizes a Markov Chain Monte Carlo method to iteratively produce parameter samples from the posterior distribution. The posterior distribution describes the uncertainty in the LPN parameters given the uncertainty in the clinical data used in tuning. These parameter samples can then be used to propagate the LPN parameter uncertainty forward to quantify confidence bounds on the 3D simulation results. Having the capability to systematically perform these functions within the same framework significantly improves repeatability of the modeling process, enables simulations on larger cohorts of patients, and produces confidence intervals on simulation results for the first time.

Automating Proteogenomic Analysis Using A Semantic Workflow Approach

Hunter Boyce

The development of tools to analyze and integrate the avalanche of heterogeneous, multi-omic data from recently developed genomics, transcriptomics, ChIP-seq, and proteomics tools, are necessary to translate “big data” into critical knowledge that enables disease prevention, detection, and personalized management. Unfortunately, the multi-omic analysis required to discover important disease mechanisms is exponentially more challenging than single-ome analysis and thus is intractable for the majority of the biological community.

To address the challenges of multi-omic analysis we are developing an open-source workflow platform: the Workflow INstance Generation and Selection MultiOmic Discovery Engine (WINGS-MDE). WINGS-MDE will enable the creation, management, distribution and effective use of multi-omic workflows. We aim to 1) create semantic workflows for performing multi-omic analysis; 2) Develop a multi-omic discovery meta-workflow engine; and 3) develop an inter-lab workflow repository that supports the annotation and dissemination of public datasets.

The WINGS-MDE system is first being applied to fully reproduce an integrated proteogenomic characterization of colorectal cancer (CRC) [1]. The original study attempted to examine important relationships across omics levels, such as which DNA mutations might be expressed at the protein level. A typical shotgun proteomics experiment compares spectra generated from a proteomics sample against a reference proteome database. However, in cancer, we expect mutations to occur. Consequently, searching a reference database may result in mis-identifying or completely missing functionally important proteins. To reproduce the original CRC study, for each patient, we first aligned raw RNASeq data, to a reference genome to call mutations. From these mutations we infered protein products and generated a novel proteome database for each patient. Building upon prior studies, we also compared the identification results from a proteogenomic (multi-omic) workflow to the typical shotgun proteomics (single-ome) workflow. These comparisons revealed that a multi-omic appraoch identified multiple functionally important mutated proteins that would have otherwise been missed.

By developing tools to reduce the barrier to integrating multiple types of heterogenous data we anticipate that it will become more possible for the broader biological community to uncover the mechanistic consequences of genetic mutations on biochemical function and accordingly the better understand the complex relationship between the genome and the functioning or malfunctioning of cells.

[1] Zhang B, Wang J, Wang X, Zhu J, Liu Q, et al. (2014) Proteogenomic characterization of human colon and rectal cancer. Nature 513: 382–7. doi:10.1038/nature13438.

BCATS 2016
www.bcats.stanford.edu

Thank you!

We would like to express our appreciation in particular to Patrick More, Mary Jeanne Oliva and Natasha Haulman for being our administrative support.

2015 Organizers

Rachel Goldfeder
Emily Mallory
Yifei Men
Kun-Hsing Yu
Weizhuang Zhou

Platinum Sponsors

Bio-X
Stanford Center for Computational, Evolutionary and Human Genetics (CEHG)
Biomedical Informatics Training Program (BMI)
Stanford Center for Genomics and Personalized Medicine (SCGPM) and Department of Genetics
Stanford Neurosciences Institute
Department of Bioengineering

Gold Supporter

Stanford Computer Forum
Genentech