Photo: Slava Gerovitch

2019 Fall PRIMES Conference

Computer Science

Computational Biology

October 20, 2019

2019 Fall PRIMES conference

Computer Science and Computational Biology

Sunday, October 20, 2019
Room 2–190, MIT

Computer Science Section

9:00 am: Welcoming Remarks

Dr. Slava Gerovitch, PRIMES Program Director
Prof. Srini Devadas, PRIMES Computer Science Section Coordinator

9:15 am: Session 1

- Ezra Gordon, “Improving round complexity of Byzantine Broadcast under dishonest majority” (mentor Jun Wan)
- Linda Chen, “Random graphs and all-to-all communication” (mentor Jun Wan)
- Sanath Govindarajan and Walden Yan, “Achieving fast fully homomorphic encryption with graph reductions” (mentor William Moses)

10:05–10:20 pm: break

10:20 am: Session 2

- Patrick Zhang, “Privacy-preserving similarity search using learned indexes” (mentor Kyle Hogan)
- Ethan Mendes, “Towards a certified defense for audio adversarial examples” (mentor Kyle Hogan)
- Andrew Shen, “Towards verifying application isolation for cryptocurrency hardware wallets” (mentor Anish Athalye)

11:05–11:15 am: break

11:15 am: Session 3

- Michael Gerovitch, “Environment-aware pedestrian trajectory prediction for autonomous driving” (mentor Dr. Igor Gilitschenski)
- Yuxuan Chen, “A deep learning approach to end-to-end autonomous driving using event-based vision” (mentors Dr. Igor Gilitschenski and Alexander Amini)
- Aditya Saligrama, “Can robust ensembling schemes improve defenses against adversarial inputs?” (mentor Guillaume Leclerc)

12:00–12:10 break

12:10 pm: Session 4

- ➢ Alek Westover, “Cache-efficient parallel partition algorithms” (mentor William Kuszmaul)
- ➢ Alex Ding, “An evaluation of UPC++ using distributed parallel graph algorithms” (mentor Dr. Yan Gu)
- ➢ Neel Bhalla, “Constructing workflow-centric traces in real time for the Hadoop File System” (mentor Prof. Raja Sambasivan, Tufts University)
- ➢ Jerry Xu, “Time – What happens if the world spins backwards?” (mentor Prof. Ari Trachtenberg and Trishita Tiwari, Boston University)

1:10–2:00 pm: lunch break

Computational Biology Section

2:00 pm: Welcoming Remarks

Prof. Leonid Mirny, PRIMES Computational Biology Section Coordinator

2:05 pm: Session 5

- ➢ Neil Chowdhury, “A method to recognize universal patterns in genome structure using Hi-C” (mentor Sameer Abraham)
- ➢ Shiv Khandelwal, “Genome-wide flame feature detection pipeline for Hi-C chromatin conformation maps” (mentor Sameer Abraham)
- ➢ Jason Yang, “The relationship between gene expression correlation and 3D genome organization” (mentors Sameer Abraham and Martin Falk)
- ➢ Vishnu Emani and Kevin Zhao, “The role of protein occupancies in DNA compartmentalization” (mentors Sameer Abraham and Martin Falk)

3:10–3:20 pm: break

3:20 pm: Session 6

- ➢ Andrew Zhang, “An explainable machine learning platform for antimicrobial resistance prediction and resistance gene identification” (mentor Prof. Gil Alterovitz and Dr. Insung Na)
- ➢ Alan Qi and Powell Zhang, “Using feature selection to identify gene significance in drug-resistant tuberculosis” (mentor Prof. Gil Alterovitz and Dr. Insung Na)
- ➢ Benjamin Chen, Neil Malur, and Hari Narayanan, “A novel framework to improve the structure of clinical trials eligibility criteria” (mentor Prof. Gil Alterovitz and Dr. Insung Na)
- ➢ Ian Balaguera, “Implementing a patient-clinician interface for biomedical templates” (mentor Prof. Gil Alterovitz and Dr. Insung Na)

4:30–4:40 pm: break

4:40 pm: Session 7

- Jonathan Yin, “Latent representations of chemical ligands to predict combinatorial receptor-ligand interactions” (mentor Dr. Hattie Chung and Michael Truell, Broad Institute)
- Sarah Chen, “Retained introns are translated and contribute antigens to the MHC I immunopeptidome” (mentors Dr. Tamara Ouspenskaia and Dr. Travis Law, Broad Institute)
- Mikhail Alperovich, “Data driven quality control for single-cell RNA sequencing analysis” (mentor Dr. Ayshwarya Subramanian, Broad Institute)

5:25 pm: End of conference

Abstracts

Session 1

Ezra Gordon, "Improving round complexity of Byzantine Broadcast under dishonest majority" (mentor Jun Wan)

Byzantine Broadcast is a well-studied consensus-building problem in computer science. A randomly chosen leader must ensure all honest users agree on the same message. Most previous literature/results for this problem rely on an honest majority of users in the protocol. It is not until very recently that Jun Wan et.al. came up with an amortized constant round protocol for a dishonest majority of users. In this project, we improve and simplify their protocol and proof. We also explore the theoretical minimum round complexity for the Byzantine Broadcast problem.

Linda Chen, "Random graphs and all-to-all communication" (mentor Jun Wan)

In all-to-all communication, all users want to exchange messages with all other users in a network. In this project, we use random graphs to represent communication networks to find ways to make communication more efficient as well as reduce cost. First, we analyze how different random graph models compare to each other regarding the graph's giant component and diameter to determine which model would be optimal in maximizing the giant component and minimizing diameter. Next, we use aggregate signatures and multisets to store the user IDs in the exchange of messages to provide improved results for the communication cost. The results of our protocol can be used in other applications as well, such as in consensus protocols.

Sanath Govindarajan and Walden Yan, "Achieving fast fully homomorphic encryption with graph reductions" (mentor William Moses)

Fully homomorphic encryption offers a way to compute on encrypted data, having an advantage over other schemes in its ability to extend to all types of data and operations. The principal concern with FHE, however, is its speed. A core reason for its lack of performance is that most applications using FHE must perform a call to an inefficient library for every binary operation (and, or, not, etc) of the computation. In this work, we present a system for automatically synthesizing efficient homomorphic versions of entire programs. This allows us to perform algebraic simplifications (e.g. not of not is identity), traditional compiler optimizations (e.g. dead code elimination), and efficient scheduling of the computation. In total, the system achieves a significant speedup over bare FHE.

Session 2

Patrick Zhang, "Privacy-preserving similarity search using learned indexes" (mentor Kyle Hogan)

Similarity search is important for online recommendation systems, finding similar songs, movies, or pieces of text for commercial users. With a large number of online websites utilizing recommendations systems such as Amazon, Spotify, and Netflix, both the users and the companies want to maintain their privacy. We construct a k-nearest neighbor search algorithm by mapping d-dimensional feature vectors onto a 1 dimensional vector through the Hilbert curve and using a learned index structure to efficiently approximate the indices of items. We then design an efficient protocol for clients to query the k-nearest neighbors of a certain item in a database that maintains privacy for both the client and server.

Ethan Mendes, "Towards a certified defense for audio adversarial examples" (mentor Kyle Hogan)

Adversarial examples are specific inputs to a neural network that result in a misclassification or an incorrect output. While most past work has focused on methods to generate adversarial examples to fool image classification networks, recently, similar attacks on automatic speech recognition systems have been explored. Due to the relative novelty of these audio adversarial examples, there exist few robust defenses for these attacks. In this talk, we examine the challenges that arise when applying defenses of adversarial examples for images to audio, as well as our preliminary findings on how to construct the first certified, mathematically proven, defense for audio adversarial examples.

Andrew Shen, "Towards verifying application isolation for cryptocurrency hardware wallets" (mentor Anish Athalye)

Proving that kernels operate securely is difficult given the complexity of modern-day kernel, yet necessary given the rise in popularity of cryptocurrency transaction software. With so many moving parts and intertwined pieces, reasoning and formulating requires large amounts of time and effort as well as attention to detail to achieve success. Microsoft's Z3 program allows us to symbolic evaluate our kernel code to verify that our kernel will operate as expected for all states that our kernel can take on, in other words, testing each possible kernel state to determine if any will make our kernel fail. Since the kernel is so complex, we choose to focus on verifying the security of launching and executing programs one at a time. In particular, we seek to prove that given a kernel with downloaded programs, each stored in memory, the launch and execution of each program do not affect the store data of the kernel or any other program. We utilize serval, a program developed to automatically lift code of any form, into code that can be symbolically executed and reasoned about in Z3, to aid in our verification. In conclusion, we make progress towards verifying program isolation for launching and running programs on a simple kernel.

Session 3

Michael Gerovitch, "Environment-aware pedestrian trajectory prediction for autonomous driving" (mentor Dr. Igor Gilitschenski)

Pedestrian safety is the primary concern when it comes to autonomous driving. There exist efficient methods for identifying static obstacles. However, predicting future trajectories of dynamic objects, such as pedestrians crossing a street, requires the development of new algorithms. We address this problem by using a deep learning approach. Using pre-annotated video footage of a crowded square and a subway station, we train a convolutional neural network to generate future trajectories of the pedestrians. We incorporate an interchangeable location bias map to account for changes in scenery, including immovable objects. This allows the network to look for consistent location-independent patterns of human movement. This technique helps increase the precision of trajectory prediction.

Yuxuan Chen, "A deep learning approach to end-to-end autonomous driving using event-based vision" (mentors Dr. Igor Gilitschenski and Alexander Amini)

End-to-end autonomous driving has recently been a popular area of study for deep learning. In this work, we evaluate the use of event cameras for the deep learned driving task. Event cameras naturally capture the motion of the scene by providing per-pixel information on intensity changes. However, processing the real-time event stream requires novel algorithms accounting for the specifics of this new sensing modality. We evaluate existing and develop new models for event cameras on steering angle prediction, and we analyze the result to compare the different methods for real-time event processing.

Aditya Saligrama, "Can robust ensembling schemes improve defenses against adversarial inputs?" (mentor Guillaume Leclerc)

A necessary characteristic for the deployment of deep learning models in real world applications is resistance to small adversarial perturbations while maintaining accuracy on non-malicious inputs. Although robust training provides models that exhibit better adversarial accuracy than standard models, there is still a significant gap in natural accuracy between robust and non-robust models which we aim to bridge. We propose a number of ensemble methods designed to mitigate this performance difference. The first, robust ensembling, reduces model-related error in the adversarial case by combining predictions from several independently trained robust models. Next, we propose composite ensembling and meta-composites of composites, which leverage features from a standard model and a robust model. Our experiments show that our proposed ensemble schemes improve both natural performance and adversarial performance, narrowing their gap under well-known PGD attacks (with default parameters).

Session 4

Alek Westover, "Cache-efficient parallel partition algorithms" (mentor William Kuszmaul)

The parallel-partition problem, which is essential to Parallel Quicksort and appears in many other algorithms, is given an array A of length n, and must partition the array based on some pivot property. The standard solution to the parallel-partition problem is out-of-place. Having an in-place algorithm is desirable because it makes the algorithm faster in practice and because sorting problems are often memory intensive so extra space may be undesirable or high cost. Kuszmaul developed an in-place algorithm for the parallel-partition problem, but the algorithm performs multiple passes over the array and thus its performance is bottlenecked by memory-bandwidth. The Blocked Strided Algorithm of Francis, Pannan, Frias, and Petit is in-place and under certain conditions performs little more than a single pass over the array. Because of this, for certain inputs the Blocked Strided Algorithm incurs very few cache misses and thus performs well. However in general this algorithm has no theoretical guarantees on span and cache-behavior. We present an in-place EREW algorithm with polylogarithmic span and provably optimal cache behavior, up to small-order factors. The resulting algorithm achieves near-ideal scaling in practice by avoiding the memory-bandwidth bottleneck. The algorithm's performance is comparable to that of the Blocked Strided Algorithm, the previous state-of-the art for parallel EREW sorting algorithms.

Alex Ding, "An evaluation of UPC++ using distributed parallel graph algorithms" (mentor Dr. Yan Gu)

With high performance computing turning toward improving parallel systems as single-processor performance reaches its bottleneck, traditional distributed parallel systems present great difficulty in their programming. As an attempt to make distributed parallel systems easier to work with, UPC++, a C++ library, exposes an API that represents the distributed memory as a contiguous global address space, similar to a shared memory system. We evaluate the convenience, scalability, and robustness of the library by testing common parallel graph algorithms that we implemented on both UPC++ and a shared memory system, OpenMP.

Neel Bhalla, "Constructing workflow-centric traces in real time for the Hadoop File System" (mentor Prof. Raja Sambasivan, Tufts University)

Diagnosing problems in large-scale distributed services is like finding a needle in a haystack. Such services can be comprised of 1000s of individual nodes. Any subset of these nodes could be involved in a given request's processing and responsible for observed problems. To help, recent work on workflow-centric tracing records the order and timing of requests' execution within and among distributed-service nodes (i.e., records their workflows). But, existing tracing systems are unable to make traces available soon after requests' execution, limiting their applicability. In this work, we demonstrate how real-time stream processing systems can be modified and used to construct traces in real time. We also demonstrate how such real-time trace construction can be used for anomaly detection.

Jerry Xu, "Time – What happens if the world spins backwards?" (mentor Prof. Ari Trachtenberg and Trishita Tiwari, Boston University)

The Network Time Protocol (NTP) is the de facto standard for synchronizing time over the internet. Its implementation can be found within almost all internet-connected computers ranging from a low-power IoT device to datacenter servers. However, as a result of the latency conscious design of the protocol, it is inherently insecure, and its built-in security protocols can be easily bypassed. It is most notably vulnerable to man in the middle attacks, many of which can be deployed easily. The exploitation of these attack can be devastating. We present this style of attack in a proof-of-concept, and explore the resulting effects on a variety of platforms, and even extending to the physical world beyond digital time. Finally, we propose methods of securing NTP to be implemented, as well as other time-sync solutions that move away from NTP and detrivialize the exploitation process.

Session 5

Neil Chowdhury, "A method to recognize universal patterns in genome structure using Hi-C" (mentor Sameer Abraham)

The expression of genes in cells is a complicated process. Expression levels of a gene are determined not only by its local neighborhood but also by more distal regions, as is the case with enhancer-promoter interactions, which can connect regions millions of bases away (Krivega and Dean 2012). The large-scale organization of DNA within the cellular nucleus plays a substantial role in gene expression and cell fate, with recent developments in biochemical assays (such as Hi-C) generating quantitative maps of the higher-order structure of DNA. The interactions captured by Hi-C have been attributed to several distinct physical processes. One of the processes is that of segregation of DNA into compartmental domains by phase separation. While the current consensus is that there are broadly two types of compartmental domains (A and B), there is some evidence for a larger number of compartmental domains (Rao et al. 2014). Here a methodology to determine the identity and number of such compartments is presented, and it is observed that there are four distinct compartments within the genome.

Shiv Khandelwal, "Genome-wide flame feature detection pipeline for Hi-C chromatin conformation maps" (mentor Sameer Abraham)

Hi-C is a genome-wide technique used to capture the three-dimensional conformation of chromatin. Constructing contact heatmaps with Hi-C data allows for the visualization of physical interactions between genomic loci. In these maps, there are recurring visual patterns that are essential to associating the spatial organization of DNA with biological mechanisms that drive phenomena like gene expression and even disease. Given that the manual demarcation of such patterns across an entire genome is extremely labor-intensive, we provide a method to computationally identify an elusive visual feature known as a flame. This feature is observed as a horizontal or vertical linear area of high relative interaction. For preliminary filtering, we adjusted for the inherent noise in Hi-C maps using Gaussian smoothing and eliminated the distance decay along diagonals using Observed Over Expected normalization. We isolated the flames using a combination of global thresholding, skeletonization, and a modified probabilistic hough-transform. Our initial detections require tuning to consistently avoid false positives, but they are promising. This pipeline will provide a new avenue to explore the biological implications of flames.

Jason Yang, "The relationship between gene expression correlation and 3D genome organization" (mentors Sameer Abraham and Martin Falk)

In some organisms such as E. coli and S. cerevisiae yeast, it is known that there is a relationship between the distance among genes and their coexpression (Pannier et. al., Kruglyak and Tang). It is also known that in general there is a relationship between gene function and genome structure (Szabo et. al). One might also expect to find a relationship between gene expression and TADs, which are domains within the genome where loci inside contact each other more frequently than loci outside. However, by analyzing data from murine brain cells, we do not find a relationship between gene pair correlation of single-cell RNA-seq gene expression and gene pair distance. Furthermore, despite the body of work linking gene expression and TAD structure, we also find no difference between gene pairs within a single TAD and between two TADs in terms of the relationship between gene pair distance and correlation.

Vishnu Emani and Kevin Zhao, "The role of protein occupancies in DNA compartmentalization" (mentors Sameer Abraham and Martin Falk)

The organization of DNA throughout the genome is a complex process to study. Analysis reveals a checker-board pattern of separation at a megabase-pair scale, called compartments, which are captured well by the largest eigenvector of the Hi-C contact matrix. The sign of the eigenvector correlates with active and repressed areas of the genome. These compartments have been characterized into two categories, called A and B compartments, which are hypothesized to be divided based upon the protein activity in the region. This project explores the factors that govern DNA compartmentalization, including the relationship between compartments and protein occupancy. In order to analyze contacts within the genome, Hi-C data was imported and the eigenvectors of the contact matrix were computed. Protein occupancy in murine cortical neurons and neural progenitor cells was measured via ChIP-Seq. Using this data, we measured the influence of several proteins on the sign of the Hi-C eigenvector via regression and Support Vector Machines (SVMs). Based on our findings, we tried to develop a simple model for compartments and explored this via simulations. We developed simple simulations of compartments based on ChIP-Seq data, and compared the results to compartments identified in experimental Hi-C maps. The results demonstrate a high correlation between the eigenvectors of the simulated and experimental Hi-C maps. In the future, different functions for attractive and repulsive forces could be utilized to model compartments more accurately. In conclusion, the computational methods are effective at determining the proteins which most significantly contribute to compartment development.

Session 6

Andrew Zhang, "An explainable machine learning platform for antimicrobial resistance prediction and resistance gene identification" (mentor Prof. Gil Alterovitz and Dr. Insung Na)

Antimicrobial resistance (AMR) threatens the effectiveness of antibiotics against bacteria worldwide, causing hundreds of thousands of deaths annually. Unsure of which antibiotics will be effective against particular strains, clinicians are in the dark on prescriptions. Traditional culture-based testing takes at least two days, during which time the patient's condition could significantly worsen. Thus, there arises a need for faster identification of a bacteria's resistances. I build a machine learning platform to determine whether a bacterial strain is resistant to an antibiotic based on its whole genome sequence data and identify the genes and mutations that cause resistance. The platform uses a Deep Convolutional Neural Network (DCNN) to quickly and accurately predict resistance, then uses a Support Vector Machine (SVM) to identify the resistance genes and mutations.

Alan Qi and Powell Zhang, "Using feature selection to identify gene significance in drug-resistant tuberculosis" (mentor Prof. Gil Alterovitz and Dr. Insung Na)

Multidrug Resistant Tuberculosis (MDR TB) is a rare form of tuberculosis which results in bacterial resistance to at least two of the most effective treatments: isoniazid and rifampin. Being able to determine if a certain strain of Tuberculosis is resistant is important for increasing patient survival so doctors do not prescribe ineffective medications. Given a binary dataset of resistant and susceptive Tuberculosis strains with mutated(1) and non mutated(0) genes, we utilized feature selection to determine the most significant genes using four models: CART, Random Forest, Genetic Learning, and Naive Bayes. After implementing the four models and analyzing the data through feature selection, we determined that gene 241a, also known as pncA, was most significant in conferring MDR TB. With previous studies indicating that gene pncA mutation contributes significantly to MDR TB, this demonstrates that our method of feature selection on binary multi feature data is successful in providing the gene of interest. With further research, our method of feature selection can be applicable to other datasets of the same nature to identify genes significant in a particular disease.

Benjamin Chen, Neil Malur, and Hari Narayanan, "A novel framework to improve the structure of clinical trials eligibility criteria" (mentor Prof. Gil Alterovitz and Dr. Insung Na)

Clinical trials often have low patient enrollment. This represents a lost opportunity for patients to receive potentially beneficial treatment, a roadblock in the approval of new pharmaceuticals, and a significant financial burden for drug manufacturers. A sizeable bottleneck in patient participation comes from the identification of trials for which a patient is eligible; eligibility criteria are highly unstructured, increasing the difficulty of queries and leading to patient frustration. The aim of this study was to build and execute a framework for increasing the structure of clinical trial eligibility criteria.

This was achieved via the development of a process to generate easily creatable and usable templates for clinical trial eligibility criteria and the implementation of Unified Medical

Language System Concept Unique Identifiers in the aforementioned templates, overall decreasing the time needed to create eligibility criteria and find potential matches among patients. The framework consisted of clustering eligibility criteria using unsupervised paradigms. This was followed by the creation of megaclusters, or generalized sets of similar criteria, change-to rules for standardizing the expression of equivalent language, and templates using a novel algorithm to build them from sets of text data. Lastly, the templates were then simplified, aggregated, and booleanized to make them machine-readable, helping to fully automate the matching process with high levels of efficiency and accuracy.

Ian Balaguera, "Implementing a patient-clinician interface for biomedical templates" (mentor Prof. Gil Alterovitz and Dr. Insung Na)

Within the Biomedical Cybernetics Laboratory run by Dr. Alterovitz, a project has been carried out by Benjamin Chen, Neil Malur, and Hari Narayanan aiming to connect clinical trials with prospective patients far more efficiently than is presently possible. Strides forward have been made, and a collection of templates were produced to standardize the input for the program. However, the templates themselves were not entirely standardized, nor did there exist and user interface via which the templates could be accessed. Hence, I have spent the last several months working to form a collection of consistently formatted templates which could be used in a sort of suggestion program acting on existing or future clinical trials. Given its universality, I decided to use JSON to reformat the templates, so I wrote a web application which incorporates a straightforward drag-and-drop interface to quickly and efficiently produce formatted templates. This interface can also import existing formatted templates (given they are in .json files), meaning they can be quickly tweaked and exported. Once this standardization was complete, I worked with Jason Zhou to make multiple interfaces through which these templates could be used. The first is a desktop and web application which consists of a simple interface where text may be entered and checked with existing templates for potential suggestions. Potential changes are underlined, and a dropdown is presented, showing the suggested correction and the template correlation to the suggestion. This makes it possible for the user to either use the quick suggestion or enter a more comprehensive pane presenting a dropdown UI making it easy to directly work with the templates. The second interface is a Microsoft Word add-in which uses the same centralized backend to provide template-based suggestions directly to a text document.

Session 7

Jonathan Yin, "Latent representations of chemical ligands to predict combinatorial receptor-ligand interactions" (mentor Dr. Hattie Chung and Michael Truell, Broad Institute)

Receptors on cellular surfaces detect and compute environmental signals. In the olfactory system, millions of small molecule ligands are sensed by hundreds of olfactory receptors (ORs) through combinatorial encoding. The ability to predict this combinatorial space of activated receptors based on the chemical structure of ligands is a powerful tool with applications to drug design. However, developing an appropriate latent representation of small molecules as inputs to training models remains a challenge. Here we aim to train a deep neural network that maps chemical ligands into a continuous latent space and extracts the structural elements relevant to receptor activation. While existing models rely heavily on lower-level syntax-based methods, we incorporate meaningful higher-level molecular features through functional group encodings alongside traditional approaches. Using this continuous molecular representation, we aim to better predict receptor activity and understand the responsible molecular features.

Sarah Chen, "Retained introns are translated and contribute antigens to the MHC I immunopeptidome" (mentors Dr. Tamara Ouspenskaia and Dr. Travis Law, Broad Institute)

Retained introns are translated and contribute antigens to the MHC I immunopeptidome
Major histocompatibility complex class I (MHC I) molecules present peptides from cytosolic proteins on the surface of cells. Cytotoxic T cells can recognize the presented antigens, and infected or cancerous cells that present non-self antigens can elicit an immune response. In this study, we identify retained intron candidates that, as a result of splicing errors, are transcribed, translated, and contribute peptides for MHC I presentation. We perform *de novo* transcriptome assembly of RNA-seq data with StringTie and identify retained intron candidates. We validate these candidates through liquid chromatography-tandem mass spectrometry preceded by MHC I immunoprecipitation in order to identify retained introns which present peptides via MHC I. Previous studies have predicted large numbers of retained introns but have been able to validate only a handful through mass spectrometry. Ribosome profiling (Ribo-seq), which provides a readout of mRNA translation, has the potential to improve retained intron predictions by distinguishing transcribed but untranslated versus translated candidates. We propose the use of a combination of RNA-seq and Ribo-seq, paired with mass spectrometry validation, to better understand and more accurately predict the contribution of retained introns to the MHC I immunopeptidome.

Mikhail Alperovich, "Data driven quality control for single-cell RNA sequencing analysis" (mentor Dr. Ayshwarya Subramanian, Broad Institute)

In recent years, single cell technologies have enabled breakthrough insights in biology by allowing us to investigate properties of individual cells. Single cell RNA sequencing (scRNAseq) profiles gene expression of individual cells, and allows us to understand cellular heterogeneity, and functions of diverse cell types. A typical scRNAseq data analysis pipeline includes multiple sequential steps, starting from quality control, feature selection, dimensionality reduction, clustering, cell type annotation, and downstream analysis to answer specific biological questions. Quality Control (QC) is an important first step where higher-quality data is retained by accounting for noise and other factors accompanying the process of data generation. In this project we surveyed commonly used QC methods, and developed alternative data-driven approaches. We benchmarked the methods on large public datasets, and evaluated their performance on ability to maximize retention of information for answering relevant downstream biological questions.