Xiaoyu Shan (Graduate student in 2nd year)

- Analysis project
- Evolution of Poly-ethylene terephthalate (PET) degrading genes
- Biological degradation of PET is of critical environmental importance and some bacteria have been revealed to hydrolyze and assimilate PET. PET is artificial synthesized only recently therefore these genes must be very newly evolved. However, the evolutionary process is still not clear.
- (Yoshida et al., 2016, Science) isolated the first bacterium degrading PET. (Danso et al., 2018, AEM) searched for putative PET degrading genes using HMM, which could be used as a database for us to start.

Sorry I am at a conference in Israel this week so have only looked into this question very quickly. I will work more on this on the returning flight.

I have a background in computational ecology and phylogenetics. Although this question is interesting, I would be also open to any other project related to comparative/evolutionary genomics.
Protein Remote Homology Detection by Sequence Embeddings Alignment

Name: Wilson Louie
Project Type: Tool Building

- Protein remote homology detection is the problem of detecting homology in cases of low sequence similarity.

- Existing approaches are typically sequence alignment based, generally from least to most effective: sequence-sequence (BLAST), profile-sequence (PSI-BLAST), profile HMM - profile HMM (HHSearch). These methods use information derived solely from sequences, whereas incorporating some structural information somehow may help with identifying homology between proteins with dissimilar sequences, but with similar structures/folds.

- Proposal: Develop (first ever?) protein sequence embeddings (akin to word embeddings in NLP) alignment approach for better remote homology detection. The learned sequence embeddings should be close together in embeddings space if their corresponding structures are similar, even if their sequences are dissimilar.
  - Part 1: Learn protein sequence embeddings from sequences with known 3D structures, using bi-directional RNNs trained on multi-structural properties prediction tasks (e.g. train ELMo architecture on predicting local structural properties + global structural properties). The resulting model can take as input arbitrary protein sequences (even those without known structures) and output sequence embedding vectors.
  - Part 2: Develop alignment program and benchmark on existing remote homology detection datasets (e.g. one curated by Hochreiter et al., 2007)
**Co-Expression Network Analysis for scRNA-seq**

- New tools like CoExpNetViz are appearing for network analysis and comparative transcriptomics.
- In this project, I would explore the output from these tools for a set of single-cell data.
- Apply several tools to the data, probably some plant single-cell expression data (e.g., here).
  - QC metrics for how well the clustering is doing.
  - Evaluation wrt orthologous genes.

---

**Orthologous Gene Network Reconstruction**

- Previously, clusters of orthologous group genes have been identified in cyanobacteria.
- Need to extend identified orthologous genes to other eukaryotic phytoplankton.
- Would explore using existing clustering tools, improve upon QC metrics, and extend to other species in the MMETSP database.

---

**Comparative Genomics for Mental Health Research**

- Either (a) comparison of the genomes of species that either do or do not exhibit evidence of mental illness or (b) comparative genomics on the genomes of patients that do or do not exhibit mental illness.
- Database of genomic information about mental health.

---

**Arianna Krinos**

IHGC example
After our discussions last week, I realized I would like to work on a project involving machine learning and disease. Particularly, disease detection or prediction seems very interesting to me. I am not bent on explicitly using machine learning for this, so I am open to other methods/algorithms/models that we have learned in class (or not!) to accomplish this goal.

Relevant papers:
- Breast-Cancer identification using HMM-fuzzy approach by Rafiul Hassan, et al.
- Machine Learning for Detection and Diagnosis of Disease by Paul Sajda
- Machine learning applications in cancer prognosis and prediction by Konstantina Kourou, et al.
Title: Fine-tuning the Epigenetic Clock

Name: Brian Xia
Type: Analysis/Tool

Description: Accumulating studies have revealed the importance of epigenetic information in modulating the aging process. In particular, Horvath established that DNA methylation levels could be used to accurately estimate an individual’s epigenetic clock (biological age). The epigenetic clock is a more precise evaluation of the aging process than chronological age and observable phenotypes given its basis in low-level processes. Considering that epigenetic modifications are reversible, the epigenetic markers underlying the epigenetic clock may be promising targets for the development of anti-aging therapies. This project seeks to further develop the epigenetic clock through the incorporation of histone modifications.

Datasets: Roadmap epigenomics

Papers: Horvath 2013; Vidaki et al. 2017; Ashapkin et al. 2017
Mutation Frequency in Different Cancer Types
Adelaide Chambers

Different cancers can be partially characterized by specific kinds of mutations, which impact their potential response to immunotherapy treatments. In addition to identifying how different cancer types differ, we can also look at how different cancer types may have similar mutation patterns, suggesting they might respond similarly to specific treatments.

• Project Type: Analysis
• Dataset: The Cancer Genome Atlas Problem
• Relevant Papers:
  • Hmeljak et. al., “Integrative Molecular Characterization of Malignant Pleural Mesothelioma”
  • Lawrence et. al., “Mutational heterogeneity in cancer and the search for new cancer genes”
Modeling Infectious Disease Spread: Ebola Case Study

Adelaide Chambers

Modeling the spread of infectious diseases helps formulate and justify public health measures like mandatory vaccinations and potential quarantines. Create an algorithm that incorporates geographic information and common travel routes to model the spread of disease under various conditions using probabilistic graphical models, which can have useful policy implications. Show this in practice in the case of the Ebola outbreak in West Africa in 2014.

- Project Type: Algorithmic
- Dataset: Humanitarian Data Exchange Ebola dataset
- Relevant Papers:
  - Xia, et. al., “Modeling the transmission dynamics of Ebola virus disease in Liberia”
  - Bindi, et. al., “Predicting epidemic evolution on contact networks from partial observations”
  - Handel, “Learning infectious disease epidemiology in a modern framework”
Predicting Response to Treatment Given A Cancer Profile
Adelaide Chambers

Personalized cancer therapies customize a cancer treatment plan according to the patient themselves. SVMs have been used to predict treatment response given gene expression information and treatment response profiles. If we use deep learning on this supervised task, we may be able to build a better predictor and potentially identify the most influential gene expression traits for predicting a treatment outcome.

• Project Type: Analysis
  • Dataset: The Cancer Genome Atlas Problem, Gene Expression Omnibus
  • Relevant Papers:
    • Huang et. al., “Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy”
Connecting epigenetic changes to transcriptional dysregulation with age

• Transcriptional noise is known to increase as organisms age. However, the source of this noise is not clear. While genetic mutations almost certainly play some role, epigenetic mutations are also implicated in transcriptional noise. Single cell datasets are ideal for researching transcriptional noise, but measurements of single cell global histone modifications are not readily comparable to single cell transcription profiles. However, by combining the single cell EpitOF data with known ChIP-Seq data, statistical estimates can be calculated for the likely contribution of epigenetic drift on transcriptional noise. By simulating transcription profiles based on our probabilistic models, we can determine how likely the observed transcriptional differences are due to histone modifications.


Christopher Rodriguez

Protective transcriptional changes in Alzheimer’s Disease

• Alzheimer’s Disease involves the down-regulation of many proteins which are prone to aggregation. Presumably, this is done to alleviate the effects of protein imbalance in late-stage pathology. However, because this effect is protective, the cells displaying down-regulation of these proteins should not be focused on even though they display significant changes when comparing early and late stage pathology. Instead, cells which do not down-regulate their proteins properly should be focused on. Transcriptional variability and genetic mosaicism are classic features of normal aging, so it is entirely possible and likely that some cells are not responding correctly to the challenges of late-stage Alzheimer’s. By searching for cells that are not displaying the correct transcriptional down regulation to alleviate the effects of protein imbalance, we can uncover potential targets for intervention.

• https://www.nature.com.libproxy.mit.edu/articles/s41586-019-1195-2
• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4855616/

Clonal Hematopoiesis and Cancer

• As we grow older, some variants of our hematopoietic stem cells expand to produce most of our blood and lymphocytes through a process known as clonal hematopoiesis. When viewed through a population genetics lens, we can describe this process as a certain (possibly beneficial) allele being fixed in our stem cell population. Cancer, the uncontrolled division of a cell, also involves certain beneficial mutations rising in frequency over time. Although there are many mutations shared between both CH and cancer, are there some mutations that are only associated with CH? The stem mutations that do well in the stem cell niche may not be the same mutations that would favor survival outside this niche, and vice versa. By studying mutations and transcriptional signatures of both increasingly monoclonal HSCs and HSC-derived cancer cells, we can determine clonal mutations that only benefit stem cell survival and proliferation.

• https://www.cell.com/cell-stem-cell/pdf/S1934-5909(18)30011-0.pdf
• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4355577/
Improving EGFR pathway model to include signaling changes due to point mutations to phosphor-Y residues on EGFR.

Analysis project to update an existing model (ODE or rule-based) so that it incorporates information about the effect mutations to residues that are phosphorylated during EGFR activation have on the downstream signaling profile. Will also improve temporal and species granularity of model. Since upregulation of EGFR drives cancerous phenotypes, this analysis could identify residues that could be targeted by drugs for most effective receptor knockout.

Using peptide MS dataset comparing mutated EGFRs to WT from my UROP in Forest White’s lab. Dataset allows for tracking of individual peptides on the timescale of seconds to minutes, which is more specific than most models allow for.

https://mct.aacrjournals.org/content/16/11/2572.abstract White paper
https://www.nature.com/articles/nbt0402-370 ODE model
Finding DNA Rearrangement Hotspots

Procedure

1. Run TCGA pan-cancer whole genomes through re-arrangement caller
2. Derive background model
3. Combine into feature space with survivorship analysis and GWAS analysis from opentargets
4. Identify novel breakpoint hotspots via ML
Finding DNA Rearrangement Hotspots

Papers

- “The Cancer Genome Atlas Pan-Cancer Analysis Project”
- “Selective and mechanistic sources of recurrent rearrangements across the cancer genome” - Beroukhim Lab
- “Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours” - Jose Tubio Lab
- “A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases” - Creighton Lab
Using VAE to find underlying structures of tumors and connecting tumors to effective treatment plans

- Similarly, we can use VAE to learn important features of tumor data with focus on advanced ones (e.g. CUP) that have poor prognosis
- Learned features can be used to find evidence of underlying primary of advanced metastasis cancers.
- With that information, we can connect patients to more specific treatment plans.
- For validation, we can perform retrospective survival analysis using in-house treatment and patient data

Simidjievski et al. “Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice”

Using Variational Autoencoder in Cancer to generate synthetic data

- The number of cancer data is usually limited in order to perform clinical assessment of its heterogeneity
- Currently used oversampling algorithms are purely based on geometry of data in Euclidean space and linear/non-linear interpolation, therefore it does not really capture biological structure of the data
- VAE (Variational Autoencoder), which effectively learns probabilistic structure of the data, can produce more biologically meaningful synthetic data for cancer.


Transfer learning in Cancer

- Cancer data nowadays exist in different forms (pathology images, sequence data, clinical notes, and etc.) and there are variances in data across different hospitals/centers.
- We would like to see if we could effectively transfer knowledge learned from one domain (images) to the other (genomics data) and consequently improve the performance of prediction
- As another example, in genomics data, there may be missing data when merging data from two different sources due to their difference in gene coverages, procedure differences, and etc.. We would like to use a statistical learning to effectively impute missing fields.

Dhruba et al. “Application of transfer learning for cancer drug sensitivity prediction”
Predicting the Outcome of Genetic Recombination using Deep Learning Algorithms

**Type of Problem:** Analytical, Tool building

**Brief Description:** "Hotspots" of recombination have been identified by multiple sources. We will attempt to use deep learning to predict the locations of these hotspots and/or identify the exact outcome of meiotic recombination.

**Dataset:** Annotations of hotspots and coldspots of the yeast genome are available, as well as yeast mtDNA used by – reference 1) as a negative control.

**References:**
https://www.nature.com/articles/srep33483
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC17209/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933199/#B14
Meta-Analysis of Controlling Nodes of Biological Networks

- Control Node is node that is heavily involved in the function and evolution of biological networks (it controls lots of nodes).
- Control nodes shift over evolutionary time.
- Can we predict propensity of gene products to become control nodes based on their sequences?
- Fairly good data for this: Avian Cartenoid network (Badyaev et al. 2019) And C. Elegans neural network (Yan et al.)
Meta-Analysis of Plant Domestication

- Hundreds of Plant Species have been domesticated by humans, most in only a few thousand years.
- Can we isolate key gene classes/events (inversions, large duplications, ploidy) that are common among domesticated species?
Build a Network Model of Human Microbiome

- Wealth of species abundance data for human microbiome across different patient samples
- Can we build a network model of different species relationships? Specifically in patients with different genetic diseases (IBD, Crohns).
- Correlation $\neq$ causality
Given the explosion of data characterizing the different molecular components that make up cells, a major challenge of modern systems biology is to organize these parts into functional pathways. We can use new computational methods to compress graphs into navigable latent spaces and thus gain insight into those pathways.

In this project, we will learn latent representations of protein-protein interaction graphs to identify critical nodes for selected cellular processes. These will be used to 1) understand the impacts of genetic polymorphisms affecting these nodes and 2) generate small molecule drugs targeting these nodes.
Leveraging DepMap CRISPR screens to classify genes as having essential hotspot and damaging point mutations

Matthew Leventhal

Problem type: Analysis, method development

Brief description: I plan to develop a classifier trained on the AVANA CRISPR screen in DepMap to identify which genes have hotspot and damaging point mutations that are more essential to cancer cell line survival than WT CRISPR knockouts. I plan to apply this classifier to identify cell lines that would potentially have essential mutations in a given gene based on the training data.

Data set: DepMap AVANA CRISPR screen 19Q3

Navigating Molecular Latent Space with Deep Learning on Protein-Protein Interaction Graphs

Neil Band and Kaitavjeet Chowdhary | Method Development and Implementation

Given the explosion of data characterizing the different molecular components that make up cells, a major challenge of modern systems biology is to organize these parts into functional pathways. We can use new computational methods to compress graphs into navigable latent spaces and thus gain insight into those pathways.

In this project, we will learn latent representations of protein-protein interaction graphs to identify critical nodes for selected cellular processes. These will be used to 1) understand the impacts of genetic polymorphisms affecting these nodes and 2) generate small molecule drugs targeting these nodes.
Name: Natasha Seelam (nseelam@mit.edu); 6.878

**Problem Statement:** Predicting evolutionary pressure of viral genomes (or proteins) from the immune system.

In the following project, we would like to model the ability of a virus to evade the immune system by rapidly mutating and evolving. To do so, we want to identify the genes or regions of a viral genome that confers immune evasion. We would like to identify sequences (genomic, but could be proteins also) that the immune system cannot natively respond to in order to develop vaccines to thwart the eventual evolutionary trajectory of a virus. A potential sub-question could be if this model framework predicts the emergence of resistance mutations.

**Problem Statement:** Identifying the genomic features that can predict viral tropism across species.

This project involves learning the features that may help illuminate how viruses may evolve and change tropism for new hosts. Some viral candidates may be Zika, Ebola, and Influenza. Some potential strategies toward exploring this problem (not limited to!) may be HMMs (modeling hidden states that could assist in recognizing whether a region is responsible for “swapping” hosts), or k-mer enrichment scores.

**Data Source:** Influenza/Ebola/etc. Virus Resource (NCBI); RVDB

**Problem Statement:** Quantifying epistasis in protein sequences to enhance directed evolution schemes.

We can think of directed evolution as an optimization problem, where the goal is to get a protein of a desired function. In this problem, we aim to understand “what regions of the protein need to co-mutate together to arrive at a better functioning protein?” (you can envision this as an initialization). Viral protein candidates may be interesting cases to explore higher-order epistasis in a protein sequence.

**Data Source:** UniProt, PDB
Prachi Sinha

Areas of Interest: Protein sequence analysis and comparison to hypothesize interactions between proteins, and effects of mutations on structure and function

Project: Tool-building, analysis
Understanding Differential Gene Expression and Regulation across Cancer Cell Lines

RNA-seq of 934 human cancer cell lines from the Cancer Cell Line Encyclopedia

RNA-Seq mRNA baseline
Organism: Homo sapiens
Reference(s): 22469905 (Filter by genes in paper)
Raw Data Provider: NHG Genomic Data Commons

Project Type: Data Analysis, Pipeline Development, Method Development
Collaborator Has: Strong Biological or Statistical Background
Contact: Sam Sledzieski - samsl@mit.edu

Gene Filtering
Select a subset of 40k+ genes to focus analysis based on information gain presence in important pathways and gene sets.

Network Learning
Generate a network connecting genes based on expression similarity.

Pathway Learning
Develop statistical model to learn direction of edges and dependence relationships.

Cross Validation
Validate network accuracy through holdout and clustering-based CV (CCV)

Cell Line Exploration
Compare networks and gene expression across cell types and disease phenotype.

GSEA: a web server for gene network construction and visualization
Gene Set Enrichment Analysis
Enhanced Gene Ranking Approaches Using Modified Trace Ratio Algorithm for Gene Expression Data

Using Machine Learning to Measure Relatedness Between Genes: A Multi-Features Model
Yun Wang, Sen Yang, Jing Zhou, Wei Du, Yanhong Liang, Cankun Wang, Fangfeng Zhou, Yuan Tien
& Qin Ma

A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models
Shayan Tabe-Bordbar, Amin Emadi, Shiht Dave Zhou & Sanush Sinha
Approximately Optimal Parallel Phylogenetic Algorithms

Idea: Exploit local structure and optimality to more quickly produce viable solutions at the cost of global optimality.

Project Type: Method Development
Collaborator Has: Strong Mathematical or Programming Background
Contact: Sam Sledzieski - samsl@mit.edu

Tree Decomposition

- Split a large phylogenetic tree into meaningful subtrees.

Algorithm Application

- Apply canonical phylogenetic approaches in parallel to each subtree.

Tree Reconstruction

- Reattach trees to find approximately optimal global solution.

Formally, a tree decomposition of $G = (V, E)$ consists of a tree $T$ and a subset $V_t \subseteq V$ associated with each node $t \in T$. We will call the subsets $V_t$ pieces of the tree decomposition. $T$ and $\{ V_t : t \in T \}$ must satisfy:

- (Node coverage) Every node of $G$ belongs to at least one piece $V_t$.
- (Edge coverage) For every edge $e$ of $G$, there is some piece $V_t$ containing both ends of $e$.
- (Coherence) Let $t_1, t_2$ and $t_3$ be three nodes of $T$ such that $t_3$ lies on the path from $t_1$ to $t_2$. Then, if a node $v$ of $G$ belongs to both $V_{t_1}$ and $V_{t_2}$, it also belongs to $V_{t_3}$.
Inferring cell type-specific changes in gene expression from changes in chromatic accessibility in disease.

Name: Sebastian Pineda

Description: Use single-cell or cell type-specific bulk ATAC-seq data to predict disease-related changes in gene expression. Possibly training a model on RNA-seq data from the same or similar cell population.

Type of problem: Analysis (maybe tool)

Dataset: Internal data.

Relevant Papers:


Idea 1: Analysis: Identifying drug targets in psychiatric disorders (e.g. schizophrenia, major depressive disorder, autism, etc.)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/#bib145
https://www.nature.com/articles/s41398-019-0451-4
https://www.nature.com/articles/s41398-019-0450-5

Idea 2: Pandemic Detection - predicting infectious disease/viral spread

https://www.nature.com/articles/s41598-019-41192-3
We would first be interested in characterizing human cancers by their differences in their chromatin accessibility, and then trying to see which variants and mutations are associated with those differences. From there, it would be interesting to see if any of those associations have predictive power.
How do the immune cell populations differ in cancer tissues, the blood, or in other reactive states? Are there genes, or chromatin states, unique to those in associated with tumors that we can modulate to improve their abilities to fight cancer? We would explore public scRNA-seq and possibly ATAC-seq datasets.

Dataset: The Cancer Genome Atlas, CancerSEA

We would like to profile the different cell variants in a tumor using publicly available single-cell and chromatin accessibility datasets in order to understand the different profiles, and understand how the mutations and variants contribute to survivorship and/or resistance.

Dataset: The Cancer Genome Atlas, CancerSEA