Tufts REU on Systems Biology and Data Science

Projects and Faculty Mentors

Machine-learning based design of antimicrobial peptides

Understanding drug tolerance of tuberculosis

Interactive Visualization and Analysis of Large Remote Data Repository

Identifying and mitigating metabolic incompatibilities in engineered cells

Intracellular bacteria

Genetically-encoded electroactive biofilms for chemiresistive sensing

Effects of colonization by the commensal fungus Candida albicans on the intestinal microbiota

Identifying bioactive gut microbiota metabolites using metabolomics

Microbial and Host Factors that Mediated Age-Related Susceptibility to Infectious Agents

Computer vision for image processing

Tumor microenvironment

Stem Cell Biomanufacturing

Discovering protein-based inhibitors of enzymes
Aeron Lab (in collaboration with Lee & Miller labs)

Machine-learning based design of antimicrobial peptides

The gut microbiota is a community of hundreds of microbial species that contribute to important physiological functions in the body. A significant perturbation to its structure, or dysbiosis, can lead to gastrointestinal disease and disorder. Increasingly, early life dysbiosis is also associated with developmental disorders and metabolic diseases. One contributing factor is the use of antibiotics, which can kill commensal bacteria along with disease-causing pathogens. For this reason, there has been intense interest in developing chemical or biological agents that can selectively kill bacteria. Antimicrobial peptides (AMPs) are an attractive class of molecules, as these peptides have been shown to target particular groups of microorganisms, in some cases at the level of individual strains. Engineering AMPs capable of selectively targeting pathogenic or other invading strains of commensal bacteria could enable new, safe ways to intervene in the microbiota to profoundly benefit human health. In recent years, computational methods have been developed to discriminate AMPs from non-AMPs and to classify AMPs into broad categories. However, there remains a need to elucidate the mechanisms underlying AMP selectivity and design rules linking an AMP’s sequence to its function, specifically to identify sequence features that confer targeting efficacy against individual bacterial species (or even strains). The goal of this project is to combine machine learning with display-based screening to design synthetic AMPs with enhanced specificity against selected bacteria.

2D embedding of 8-dimensional molecular descriptor data generated by modlAMP and visualized in MATLAB using t-SNE.
Understanding drug tolerance of tuberculosis

Tuberculosis is caused by infection with *Mycobacterium tuberculosis*, and treatment of TB remains lengthy because some bacteria are tolerant of antibiotic treatment. Mycobacteria grow asymmetrically, giving rise to differences in growth behaviors and antibiotic tolerances. We aim to characterize what is different about mycobacteria that are drug tolerant from those that are killed quickly by antibiotics with the long-term goal of optimizing treatment strategies. Our work combines experimental and computational approaches, and each student project engages both types of work. We use live-cell microscopy, microfluidics, and fluorescent reporter strains of mycobacteria to observe cell growth and drug response. We quantify the contribution of each cellular factor to drug response by building and analyzing computational models.
The goal of this project is to support the visualization and exploration of the large amounts of biological and chemical data collected by the other research efforts in this proposed REU activities. This project will have two interconnected branches of investigations: (a) the development of useful interactive visualizations that are tailored for the collected biological and chemical data; and (b) the research of advanced data structures and techniques that will enable visualizations at scale to handle very large amounts of data stored in data centers and repositories. Visualizations are often the first artifact created by scientists and engineers to visually verify and understand their newly collected data. When combined with high degrees of interactivity, researchers can quickly explore the data to confirm hypotheses about the data, as well as discover unexpected phenomenon or patterns. However, most common visualizations (e.g. in Excel, MATLAB, or R) are limited in interactivity and cannot easily scale to visualize $10^6+$ data points.

In order to create effective large-scale interactive data visualizations, developers often need to create custom-made visualization systems that exploit the properties of the data, users, and tasks. The Chang research group will work with REU students to design and develop custom visualizations in support of the other research activities in this REU proposal. This process will teach the REU students in communicating with the end users to understand their needs. More importantly, the students will learn to convert such requirements into system functional specifications that are incorporated into the final design. Finally, the students will develop these visualizations using state-of-the-art languages and platforms such as Javascript (D3), Java, Processing C++, and OpenGL (or WebGL).

In addition to the practical aspect of developing visualization systems, the students will also participate in researching effective algorithms and techniques for supporting interactive visualization and exploration with large-scale remote data repositories (e.g. databases of annotated genomes) that are of increasing importance in biological research. As data sizes increase, data is often stored in databases or data centers that are run remotely (e.g. in the cloud). Integrating visualizations with these remote data repositories adds to the complexity of the overall system due to the overhead incurred from communicating with these data repositories and transferring the data over the networks. To address the challenge of speed in data delivery, we will investigate asynchronous pre-fetching and caching mechanisms based on the user’s expectations about the user interface elements. This concept is a generalization of pre-fetching from the computer graphics literature, but extends the degree-of-freedom from six in standard 3D graphics to an arbitrary number that is more typical of exploratory visual analytical system.
Hassoun and Nair Labs

Identifying and mitigating metabolic incompatibilities in engineered cells

Increasing understanding of metabolic and regulatory networks underlying microbial physiology have enabled engineering of progressively more complex cells for production of biofuels, bioplastics, fragrances, and drug molecules. Despite the best efforts by engineers to describe various biological systems from a reductionist perspective, unpredictable outcomes still emerge from unforeseen interactions between heterologous synthesis pathways and the host cell. Such interactions are difficult to predict when designing engineered cells and often manifest during experimental testing as inefficiencies that need to be overcome. Despite advances in tools and methodologies for strain engineering, there currently remains a lack of tools that can systematically and quantitatively identify such interactions.

The objective of this project is to investigate a fundamental yet overlooked interaction mechanism that presents itself at the metabolic level – a phenomenon we dub as “metabolic network disruption”. The premise behind this phenomenon is that high concentrations of heterologous enzymes and metabolites lead to formation of unintended byproducts that disrupt the host metabolic network and decrease biosynthetic pathway productivity. Such disruption is observed experimentally in many studies where careful gene knockout strategies consistently fail to completely silence byproduct formation. Our hypothesis is that substrate promiscuity, where an enzyme transforms other substrates in addition to its natural substrate, is the culprit behind such unexplained byproducts. In this summer project, we will test this hypothesis by developing tools and workflows to computationally identify, quantify, and provide interventions to overcome such metabolic disruptions. This work will contribute to a bigger project that involves the experimental validation of our computational findings.

Metabolic cellular network (green oval) consisting of native metabolites and reactions, and heterologous synthesis pathway (red) consisting of heterologous metabolites and reactions. Heterologous enzymes catalyze their own natural substrates. In some cases, they promiscuously catalyze native host metabolites. The resulting products may be new metabolites that are not part of the metabolic models or existing metabolites. Native enzymes act on their own natural substrates and in some cases on heterologous metabolites, resulting either in native metabolites or new metabolic products.
The Isberg laboratory focuses on determining how intracellular bacteria hijack host cell secretory processes to allow formation of a replication compartment. The model system that will be studied by the REU students is the growth of the bacterium *Legionella pneumophila* within either *Drosophila* cells or the amoeba *Dictyostelium discoideum*. Formation of the replication compartments in these cells requires a specialized bacterial protein translocation system that moves at least 275 different proteins into host cells. Although the identity of these proteins is well known, one of the most persistent problems in their analysis is that it is extremely difficult to identify the minimum protein set necessary for intracellular growth. To solve this problem, we have developed a multiplex system called iMAD, in which the phenotypes of more than $10^4$ bacterial insertions can be followed individually in host cells depleted for individual proteins. In our current version of this strategy, the relative fitness of each mutant is determined by high-density sequencing, using abundance of reads as a measure of fitness for each insertion mutant. The data are then analyzed by hierarchical clustering, with each phenotypic cluster of genes assigned to a single functional group. We have shown that combining mutations from different functional groups allows us to construct strains that are defective for intracellular growth, allowing a partial solution to genetic redundancy. Remarkably, proteins that define a single hierarchical cluster appear to perform identical functions in support of intracellular growth. The REU student will take advantage of this strategy, and will first perform bioinformatics analysis of iMAD datasets that they will be provided from data already accumulated in the laboratory. Based on the pathway predictions that the student makes from these large datasets of up to 400,000 phenotypic characterizations, the student will identify pairs of mutants expected to give defects in intracellular growth. The student will then use these predictions to construct double mutants and test the hypothesis that simultaneous loss of these two proteins will result in defective intracellular growth in *Drosophila* cells and *D. discoideum.*
The overall objective of this proposed research is to design and develop biologically derived, electrically conductive extracellular matrices (ECM) as a new class of electronic transducers in chemical sensor development. In particular, we will exploit exoelectrogenic bacteria that are widely available from deep marine sediments as cellular factories to generate metalloprotein-based ECMs, whose electrical resistance can be effectively modulated through the specific interaction with target analytes. The initial project will involve the development of ECM chemiresistor arrays by initiating electroactive biofilm growth on interdigitated microwires, as well as proof-of-concept experiments and computational modeling to investigate and optimize their performance against a wide range of molecular analytes. Depending on the results, next-phase studies will target the modulation of sensor specificity by chemically modifying the extracellular electron pathways or genetically engineering the matrix conductive properties with synthetic biology tools (through collaboration with Nair Lab).
Effects of colonization by the commensal fungus Candida albicans on the intestinal microbiota

Our research focuses on interactions between the opportunistic fungal pathogen, Candida albicans, and its environment. This organism is a common commensal of the human GI tract and, in this environment, interacts with the bacterial microbiota, components of the host immune system and the metabolites that make up the intestinal milieu. While colonization by C. albicans in a healthy host is benign, organisms colonizing the GI tract represent the reservoir from which C. albicans may spread to different parts of the body. In immunocompromised patients, spread by C. albicans can result in fatal candidemia. Higher levels of GI colonization are associated with higher rates of candidemia and, thus, the factors that influence levels of C. albicans colonization are important to elucidate. To understand how C. albicans influences the gut environment, our research employs a murine model. Mice are not naturally colonized by C. albicans and we can therefore study the effects of this organism by comparing mice that have been experimentally colonized with C. albicans and mice that have not. We and others have demonstrated that colonization with C. albicans results in changes in the composition of the bacterial microbiota. We are conducting further studies to understand the biological significance of the changes that have been detected. As part of this project, students will use the murine model to obtain samples, extract DNA and characterize the bacterial groups that are present, and study the composition of the bacterial microbiota using data obtained through Illumina sequencing. Standard data analysis using established methods will be conducted. There will also be opportunities to develop novel data analysis methods specific for the needs of this project.
Identifying bioactive gut microbiota metabolites using metabolomics

The project’s overall goal is to identify bioactive small molecules that are derived as a result of microbial metabolism in the mammalian digestive tract and can modulate inflammation in mammalian tissues. The student researcher will participate in both experimental or computational aspects of the project. The experiments will utilize mass spectrometry methods to broadly characterize the products of microbial metabolism in various tissue samples, including intestine, liver, and brain. The computational aspect of the project deals with developing models and software tools that use information from publicly available genome database to predict the microbiota metabolites, and determine the biochemical pathways and microbial organisms responsible for these metabolites.
The enhanced susceptibility of elderly individuals to infectious disease, i.e., “immunosenescence”, is multifactorial, encompassing host and microbial genetic determinants, as well as environmental factors such as host nutritional status. *Streptococcus pneumoniae* can cause life-threatening pulmonary and systemic infection, particularly in the elderly. The laboratory of Dr. Meydani (and others) has shown that T cell-dependent immune functions in elderly are compromised and can be partially mitigated by Vitamin E dietary supplementation. Dr. Camilli has generated methods for genome-wide determination of all microbial genes required for colonization and growth within the mammalian host [23, 24], and utilized this approach to identify host immunodeficiencies that lead to enhanced susceptibility to *S. pneumoniae* in a murine model of sickle cell anemia. The proposed project will address the sources of immune-senescence in a murine model of *S. pneumoniae* pulmonary disease, and determine if this is due in part to vitamin E-reversible immune defects. To this end, we will use high-throughput screens of large banks of *S. pneumoniae* transposon mutants to generate “immuno-fingerprints,” consisting of the set of bacterial genes required for growth in a mouse. The contribution of *S. pneumoniae* genes will be determined by measuring the change in relative frequency of transposon insertion mutants via massively parallel sequencing of the transposon junctions before and after infection. Genes with significant roles in infection will be identified in a one-sample t-test with Bonferroni correction for multiple testing. The REU student will perform the screens in young and aged mice, or in aged mice supplemented with Vitamin E. Analysis of the *S. pneumoniae* immuno-fingerprints for each of these three hosts will provide insight into (a) the identity of *S. pneumoniae* genes required for growth and colonization in the mammalian host, and (b) how these genes interact with host immune components that are modulated by age or diet. An REU student performing this project would gain hands-on experience in performing a genome-wide genetic screen and in statistical analysis of the resulting data.
Computer vision for image processing

The goal of this project is the development of computer vision processing methods for automating the analysis of high throughput microscopy data to address problems such as identification of lipid droplets in fat and liver cells and quantification of their size and shape given images and/or video streams from one or more microscopes. The project will involve close collaboration with faculty and students from the chemical and life sciences. The student researcher is expected to have prior programming experience with Matlab, python, C/C++ or a similar computational tool or language. The initial part of the summer will focus on developing the necessary background in image processing and analysis while the bulk of the times will be devoted to applying the methods to a real-world problem.

(A) Original image (B) Result of Random Forest Classifier (C) Final image showing individual droplets
Oudin Lab

Tumor microenvironment

The goal of the Oudin Lab is to investigate the role of the tumor microenvironment in cancer metastasis and drug resistance. Metastasis, the dissemination of cells from the primary tumor, is the leading cause of death in cancer patients. Increasing our understanding of the mechanisms driving metastasis will help identify biomarkers that can be used to predict it and characterize targetable signaling pathways to that can be used to prevent it. The tumor microenvironment composed of immune cells, blood vessels, nerves and the extracellular matrix plays a critical in driving tumor progression and metastasis. We use microfluidic devices to model these environments in vitro, and study the effect of biophysical, chemical and electrical cues on cancer cell migration and ultimately metastasis. Our goal is to take a systems approach to understanding how multiple directional cues cooperate to drive invasion and how tumor cells integrate these multiple signals using microfluidics and immunocytochemistry.
Stem cells hold great promise for the engineering of tissues to replace or reconstitute native tissue after injury or disease. Scalable processes for the expansion of stem cells (SCs) and their differentiation to therapeutically useful phenotypes according to firm specifications are highly desirable. Given the prevailing empirical approaches to SC differentiation in dishes, there is little insight about the underlying biological determinants and their practical manipulation in commercially relevant, large-scale systems. Quantitative models are essential for the rational design and prediction of SC differentiation processes and their translation to biomanufacturing practices. The ‘average’ population behavior described by current models is unsuitable for the inherently heterogeneous SC ensembles. To that end, our laboratory works on quantitative models of SC specification, namely, population balance equations (PBEs) coupled to measurable distributions of cellular traits. We utilize genome-editing methods such as CRISPR/Cas9 to engineer SC lines suitable for lineage tracing experiments. These cells are expanded in stirred-suspension bioreactors and subsequently subjected to directed differentiation along particular lineages such as pancreatic endoderm and cardiac mesoderm. The physiological state functions of the cell populations are extracted through a combination of experiments and computational solution of the relevant inverse PBE problem. This leads to a quantifiable landscape of commitment along domains of variables such as differentiation stimulus concentrations, duration of treatment and bioreactor variables such as agitation, cell density etc. The outcomes can be employed for fine-tuning the yield of therapeutically useful SC-derived progeny. As part of the planned activities, undergraduates will take part in the design of experiments to test and optimize combinations of differentiation factors, their concentrations, and duration of treatment of stem cell-derived pancreatic or cardiac cell precursors. In addition to routine experimental method (e.g. qPCR, immunostaining) undergraduate students in the program write codes for numerical calculations based on the models developed in the laboratory.
Genetic code manipulation is a powerful approach for discovering new therapeutics and understanding basic biology. However, cells evolved to use the conventional genetic code, not an expanded genetic code. How can cells be engineered to efficiently utilize an expanded genetic code? This project will explore this broad question in yeast using a combination of rational design, screening, and data analysis.