Semester-long Projects for Bioinformatics class

1. Your own project (ex: related to your job, academic interests)
2. Association between germline SNPs and cancer in African Americans
   (for presentation at University of North Texas in April)
   Patricia Francis-Lyon, USF; Bob Horton, Senior Data scientist at Microsoft
   a) Genomic data analysis (4) have team members, need a System Administrator for GCP
   b) Proteomic investigation using homology modeling. (1-2)
3. Django version of Active/Transfer Learning with Medical Imaging
   Bob Horton, Senior Data scientist at Microsoft
4. Epigenetic Control of DAX-1 Expression: Is there a master regulator?
   Christina Tzagarakis-Foster, Associate Dean at USF
5. Predicting Unplanned Hospital Readmission Using Natural Language Processing
   of MIMICIII Discharge Notes
   Sue Sarafrazi, Molecular Diagnostics Manager at Ultragenyx (2-4)
   Could have a team lead who has taken NLP and wants to delve further
6. Recurrrence Networks to model EEG data using pyunicorn
   Bill Bosl, USF, Boston Children's Hospital
7. Lactose intolerance Genomic Investigation:
   EDA and visualization producing Shiny web app. (Up to 3)
   Lu Chen, USF
8. Data Collection& Analysisi App for Birth Clininc in Malawi
   Meera Nosak
9. Open Source project motivated by Google Summer of Code (2-4)
10. Investigation of Genomic dataset with PLINK - gatk project (gatk covered in class, PLINK self-study)(2-4)
11. TCGA Biolinks Genomics and epigenomics (self study with scripts from previous Breast cancer team)(2-3)
12. Medical imaging w/public data self study (3-4)
13. Cracking the "Sepsis" Code: Assessing Time Series Nature of EHR Data,
    and Using Deep Learning for Early Sepsis Prediction (2-4)
    (study CNN, LSTM, XGBoost scripts produced for challenge)

2. Association between germline SNPs and cancer in African Americans (1 who can serve as Sys Admin)

a) African Americans have been shown to have higher incidence, higher mortality, and more aggressive forms of some cancers. However, little is known about a genetic basis for this disparity. Detecting associated SNPs could lead to improved risk detection and outcomes for African Americans. For example, Hilliard has pointed to the importance of investigating genetic basis, focusing on a variant of the TRPV6 calcium ion channel gene that is present in ~50% of African Americans but is not found in other ethnicities. TRPV6 has been shown to be up-regulated in cancers including breast and prostate, and to be correlated with progression. Hilliard suggests that African Americans who carry this variant are maladapted to the high calcium food environment in the US, contributing to higher rates of more aggressive cancers. Finding a genetic basis for breast and prostate cancer disparity could help spark research into actionable environmental factors, possibly resulting in improved outcomes for African Americans.

This study will investigate germline SNPs, including TRPV6, for association with the higher incidence and more aggressive behavior of cancers in African Americans. Exploratory data analysis and classification of case vs. control will be performed. Classification models will be developed that are highly interpretable in addition to models with high predictive accuracy. Models will include regression, elastic net, decision tree, random forest, support vector machine, XGBoost and artificial neural network. After separate analysis, dbGaP datasets will be combined to achieve greater power in detecting genetic associations, including interactions. Data requested in this proposal may be combined with non-dbGap data, such as TCGA, but only datasets to which access has been granted that are de-identified and IRB approved. We do not believe this creates any additional risk to participants. Findings will be shared with the scientific community via publication in peer-reviewed journals and presentation at conferences. Software for classification models will be disseminated as open access software. We will comply with all Use Restrictions of the requested datasets.

b) Proteomics:

The SNP of the TRPV6 haplotype that is implicated as the carrier of positive selection is M681T. An issue is that this region is very difficult to resolve with crystallography. (I believe this is stated in one or more of the 3 journal articles you read). Therefore it's difficult to inspect a comparison between the ancestral protein related to rat and other mammals) and the derived, which is broadly found in humans.

Homology modeling: I'd like to start with a mammalian TRPV6 that aligns fairly closely to human, and place the derived amino acids on the backbone utilizing Modeller or some other software to perform the energy minimization. I began to look into this almost a year ago, and identified some protein structures that might be useful for homology modeling, will provide these.

3. Django version of Active/Transfer Learning with Medical Imaging (2-4)

Utilize Django to create a web interface to an image labelling system
Start with R/Sniny code by Bob's Active/Transfer Learning team last year
Transfer and Active Learning applied to Medical imaging

If possible, extend:
* make it a more general labeling tool (e.g., support bounding boxes, maybe arbitrary shapes for semantic segmentation)
* design it so we can swap in more sophisticated (or specialized) example selection approaches
* support multiple labels (so we can experiment with multi-task classifiers for learning interpretable representations).

4. Epigenetic Control of DAX-1 Expression: Is there a master regulator?

Introduction:

My research at USF is focused on the orphan NHR, DAX-1. The name DAX-1 comes from **D**osage Sensitive Sex Reversal, **A**drenal Hypoplasia Congenita, on the **X** chromosome, gene **1**. Mutations within the DAX-1 gene lead to a disease called Adrenal Hypoplasia Congenita (AHC). AHC patients fail to develop a normal adrenal gland, which ultimately is lethal due to the lack of production of critical adrenal hormones. Furthermore, mutations leading to the duplication of the region of the X-chromosome expressing the DAX-1 gene result in the reversal of a male phenotype to a female phenotype. Patients with this duplication (called Dosage Sensitive Sex Reversal) are genetically male but present with female characteristics. This suggests that DAX-1 at least in part, regulates the intricate process of sex determination very early on in embryonic development and that a precise amount of DAX-1 gene expression is required to coordinate this event properly.

DAX-1 has been shown to interact with other members of the nuclear hormone receptor family such as Estrogen Receptor (ER), Androgen Receptor (AR), Thyroid Hormone Receptor (TR) and Steroidogenic Factor (SF-1). The result of this interaction ultimately ends in transcriptional repression – the block or shutting down of gene expression. Research aimed at elucidating the exact genes that are targeted and repressed by DAX-1 and understanding the molecular mechanism of DAX-1 repression is still in the early stages. However, given the outcome of DAX-1 mutations or DAX-1 gene duplication in mediating disease phenotypes, identification of DAX-1 transcriptional targets is expected to reveal key genes in a variety of developmental cascades.

Epigenetic Control of DAX-1 Expression

Outside of its role in human development in the embryo, DAX-1 is known to be expressed in adult human tissues such as the testes, ovaries, adrenal, pituitary, and hypothalamus. Examination of normal human breast tissues (both cell lines as well as patient tissue) reveals that DAX-1 is expressed in low, but detectable amounts. We and others have found that DAX-1 expression is significantly reduced in several ER-positive

breast cancer cell lines. This finding is fascinating as it has been proposed that DAX-1 may have tumor suppressive properties and loss of DAX-1 expression may be one of the factors leading to the formation of tumors in human tissue such as the breast. The question arises, then, how does DAX-1 switch from being expressed in normal tissues to nearly undetectable levels in cancer cells? ***Is there a master regulator of DAX-1 expression?***

In the past 20 years, much research has focused on elucidating the different mechanisms that control gene expression. During this time, the field of "epigenetics" exploded. The word "epigenetic" literally means "in addition to changes in genetic sequence." The term has evolved to include any process that alters gene activity without changing the DNA sequence and leads to modifications that can be transmitted to daughter cells.

Many types of epigenetic processes have been identified—they include methylation, acetylation, phosphorylation, ubiquitylation, and SUMOlyation. Epigenetic processes are natural and essential to many organism functions, but if they occur improperly, there can be major adverse health and behavioral effects. Perhaps the best-known epigenetic process, in part because it has been the easiest to study with existing technology, is DNA methylation. This is the addition or removal of a methyl group (CH_3), in regions of the genome called CpG islands. CpG islands are made up of long stretches, anywhere between 500 and 2000 nucleotides, of the DNA nucleotides "C" (cytosine) and "G" (guanine). Typically, these are located in the promoter region of a gene. DNA methylation was first confirmed to occur in human cancer in 1983. Since this time, DNA methylation has been linked to many other illnesses and health conditions.

Using several different methodologies in my research lab, we have confirmed there is a correlation between DAX-1 expression in cells (both normal and cancer cells) and methylation status. Furthermore, employing a variety of protein-based analysis aimed at identifying key epigenetic factors, we have begun to elucidate a mechanism for the epigenetic disparity of DAX-1 between healthy and cancerous human cells.

5. Predicting Unplanned Hospital Readmission Using Natural Language Processing of MIMICIII Discharge Notes

Drs' notes represent a vast wealth of knowledge and insight that can be utilized for predictive models using Natural Language Processing (NLP) to improve patient care and hospital workflow. In this project we want to predict hospital readmission with discharge summaries.

After completing this project, you will learn

- How to prepare data for a machine learning project
- How to preprocess the unstructured notes
- How to build a simple predictive model
- How to use techniques such as bag-of-words and more sophisticated models such as LSTM, CNN, ...
- How to assess the quality of your model
- How to decide the next step for improving the model

6. Recurrrence Networks to model EEG data using pyunicorn

Challenge in time-series, grasping a new algorithm, and software engineering.
Implementing code using the pyunicorn package:
http://www.pik-potsdam.de/~donges/pyunicorn/.

I have some slightly messy code now that takes as input a file with EEG time series in a format called European Data Format or .edf files. It's a commonly used open binary format for EEG files. It computes several nonlinear measures from each of the time series in the file. I'd like to expand the capability of this to:

(1) Compute several new nonlinear measures derived from recurrence networks
(2) possibly implement joint recurrence networks, which involve pairwise interaction between different sensors or time series.

Both of these may require some background research because I don't even know all the measures that should be derived. There may also be some experimental work to derive 3 parameters needed in the calculations. I would be pleased also just to have the code cleaned up a bit and generalized to handle these new calculations. In the near future this will become part of a larger processing pipeline and database system at Boston Children's Hospital that will be used for many research projects by me and colleagues at Harvard Medical School.

Besides the link above, there are a couple of papers on recurrence networks. One paper is nearly 80 pages and may be too much. I shorter introduction is attached.

Article by some of the same authors:
https://arxiv.org/pdf/0908.3447.pdf

I can go ahead and pitch this project tomorrow and see if we have students who welcome a challenge in time-series, grasping a new algorithm, and software engineering.

7. Lactose intolerance Genomic Investigation:
EDA and visualization producing Shiny web app. (Up to 3)
Lu Chen, USF

The aim of this project is to investigate the lactose intolerance dataset, which is a genomic dataset from OpenSNP. We will perform exploratory data analysis (EDA) and visualization in R, and advance to producing interactive web apps utilizing Shiny. Shiny web apps are used in industry to host standalone apps on a webpage or embed them in R Markdown documents or build dashboards. These permit staff scientists to interact with data, enhancing understanding and problem solving. If time permits and there is interest, we can explore some further visualization techniques using D3.

8. Data Collection& Analysis App for Birth Clinic in Malawi
   Tailor open source code by Medic Mobile to collect data, perform data validation and if time some visualization.

The Community Health Toolkit

VISIT THE CHT WEBSITE

Medic Mobile serves as the technical steward for the Community Health Toolkit (CHT) open source project. The CHT provides you with resources to design, build, and deploy digital tools for community health. It includes open source software frameworks and applications, guides to help you design and use them, and an active community for creation, collaboration and support.

Get Involved

Are You A Developer?

Read the documentation and dive in to the developer sandbox
Join The CHT Slack channel for developers
Review "Help Wanted" tags in Github
Propose or build an integration

9. Open Source project motivated by Google Summer of Code (2-4)

Look through the projects that were funded in 2019 and 2018 to see if there's an open source code base you would like to get involved with.

Propose a project to a sponsoring organization, for you to work as a paid Google Summer of Code intern.

Recent sponsoring organizations include:

OpenMRS, cBioPortal for Cancer Genomics, Global Alliance for Genomics and Health, InterMine, LibreHealth: Healthcare for Humanity, Open Bioinformatics Foundation, Open Data Kit, Genes, Genomes and Variation, CHAOSS