# Google Summer of Code 2019 Proposal

Google
Summer of Code

THE UNIVERSITY OF
TENNESSEE
HEALTH SCIENCE CENTER

**Image source: Google images**

# Table of contents

# Contact Information

| | |
|---|---|
| Name | Shayantan Banerjee |
| Country | India |
| School and Degree | Indian Institute of Technology Madras<br>Master of Science in Computational Biology<br>Expected to complete by December 2019 |
| Email | banerjeeshayantan@gmail.com |
| Phone | +91-8017474272 |
| Preferred communication | Email, Video conference |

# Brief Intro

Let me first introduce myself to all of you. I am **Shayantan Banerjee** from Indian Institute of Technology Madras. I am pursuing my master's degree here under the supervision of Dr. Karthik Raman and Dr. Balaraman Ravindran. I am trying to develop machine learning models to understand the context of mutations in a cancer genome. Primarily I work with TCGA data and use state of the art bioinformatics tools to collect, analyze and interpret cancer mutation data.

I did my Bachelors in Computer Science from the West Bengal University of Technology, Kolkata. After a couple of research internships and based on my supervisor's advice, I understood the full potential of computational biology and the extent to which I can contribute to this field. During my first research internship at the National Centre for Radio Astrophysics Pune, I was exposed to astronomical big data for the first time. After a fruitful couple of months of intensive data analysis that included predicting the photometric redshifts of different astronomical sources and analyzing star clusters, the internship came to an end. On the last day, I expressed my satisfaction in working with such enormous datasets for the first time in my life and my supervisor suggested me to look into biomedical data analysis, where the next boom of the so called "big data revolution" is happening.  I decided to give it a try. On my final undergrad year, I took up an intern position at the Indian Statistical Institute Kolkata at the Centre for Soft Computing Research. The project I was allotted could be roughly summed up as reproducing and extending the results from this paper. The primary goal was to try a host of classification algorithms for the identification of differentially expressed genes from microarray data. (More details included in the Experience section).

As a person I am enthusiastic about independent research and love to learn newer programming challenges in biomedical engineering. Due to the same reason, I applied for this particular org in this summer, as it showcased a challenging topic that had a considerable research component to it.

# Experience

- **Microarray data analysis**
    - Downloaded eight different datasets from the caGEDA website maintained by the University of Pittsburgh. These were eight different types of cancer (tumor-normal matched ) data.
    - Studied different transformation and normalization methods to improve signal-noise ratio. Some of them included shifting the data by a number of medians of the channel followed by global mean adjustment to make all the measurements have the same overall mean and finally log-transformation.
    - Different classifiers were trained on each of these datasets including SVM, Random forest, Logistic regression, Perceptron and Decision Trees.
    - Feature selection was performed before classification and 10-fold cross validated accuracy was reported to see if the model is generalizable.

- **R/Bioconductor**
    - I have worked with several R based microarray data analysis packages. Some of them include the **"affy"** package to work with affymetrix oligo arrays. **"Gcrma"** to normalize the data and extract the probe-set intensities and plotting them using **boxplot()** to see whether they have comparable median intensities.
    - Genes differentially expressed between the tumour-normal samples using the **"limma"** package. These include fitting a linear model to each gene and getting p-value for comparisons.
    - Finally getting the list of top 'n' differentially expressed genes after doing multiple hypothesis testing.

- **Python**

  Last year I created SmartEHR, a web-based tool to analyze free text clinical notes using Natural language processing and Machine learning techniques. This project was among the top 20 selected ideas at the 24hr GE healthcare precision health challenge hackathon. It was also selected as the best project at the HSBC Shaastra AI challenge 2019. This tool utilized NLP techniques to mine biomedical big data in the form of free text clinical notes from the MIMIC III database. After the required text preprocessing, standard machine learning algorithms were applied to perform three major prediction tasks.
    - 30-day hospital readmission chances
    - Disease diagnoses codes (ICD9) prediction
    - Top ten keywords specific to each clinical note

  A web-framework was also created using the Python flask environment for ease-of-use. Below are some of the tools used for the purpose.
    - I have good working knowledge in **NLTK, scikit-learn, gensim** for machine learning purposes and matplotlib for visualization. (Code on github).
    - I have worked with the **python flask** environment and have created flask apps for the above project

➢ **Next generation sequencing techniques**
I have an intermediate level experience working in the following areas:
- Command line based tools to analyze genomic big data, downloading data from **TCGA, COSMIC, ENSEMBL** etc & handling different data types(.bam, .sam, .bed, .vcf  .fastq etc)
- Preprocessing raw bam/fastq files and producing quality charts for analysis.
- Alignment tools like **SAMTOOLS, BWA-MEM, BOWTIE2** etc
- Variant calling tools (mostly cancer-specific) like **Varscan, GATK, Mutect2** etc
- Finally, analyzing the variants called with gold standard datasets

## Programming strengths and preferences

➢ Primarily I code in **Python** and use **scikit-learn** extensively for my work. Apart from that I work with the two basic libraries of Python (**Numpy and Scipy**) to implement tasks like transforming data, feature selection and ensemble methods using just a few lines.

➢ My next language of interest and expertise is **R**. If i want to do some quick and dirty analysis, I would very much prefer R. Starting from reading data from databases (I have used **RMySQL, RPostgresSQL, RSQLite, xlsx**) to manipulating data (I have used dplyr and tidyr) and finally my favourite part : visualization (I have used ggplot2, ggvis), R has always been my first choice.

➢ I have somewhat moderate experience in building web applications using the **Python flask environment.**

➢ In my undergrad years I have used C/C++ extensively. One of my favourite pastimes included trying to solve competitive coding problems using the above two languages.

## Why GSOC with UTHSC?

➢ **Motivation:**
I have always been enthusiastic about projects related to biomedical informatics. The main reason for this is because it's an interdisciplinary field and I can contribute significantly to improving patient conditions in a clinical setting. I am also very much interested in **translational research.** There are 1000s of diseases that are incurable and I think it's high time for biologists to pair up with clinicians, bioinformaticians and policy makers to answer complex medical questions. **One small example** of which I found in my research: Since I am working on making tools to prioritize cancer-causing mutations (drivers and passengers), one major part of my work involves tuning machine learning algorithms to get better predictions than benchmarking studies. But after I have found the best model, my job would be to look at those individual features that contributed the most to my model and assign biological sense to it. This can be in the form of biological pathways, drug combinations already known to affect those pathways or finding clinical relevance of those features. There can't be a better way to spend my summer than working with real patient data  and analyze them to help a clinician or a pharmacist make better decisions.

- **Prior exposure the biology and informatics:**
    - One semester course in Introductory cellular, molecular biology and genetic engineering.
    - One semester course in Computational systems biology
    - One semester course in Cancer biology
    - One semester course in Introduction to machine learning
    - Two semester course in Linear algebra and Probability

I have worked on clinical data to make robust text processing models to analyze free text notes. I have also worked on time series data from stock markets, but I have never worked on real time sensor data from patients. That's one major reason for taking up the project. You have your time series patient data. You also have your gene expression data from a cohort. Is it possible to combine the two? If so, the model we come up with, is it generalizable or very specific to the set of patients we are studying? How do you predict gene expression data at regular time intervals when you don't have training data to start with? And finally after integrating the two, does it improve prediction performance considerably? There questions are very exciting to explore further and I intend to do so by working with UTHSC this summer.

## Regarding commitments this summer

These are the clashes of GSOC timeline with my academic commitments

| Time period | Phase | Reason | Time commitment |
|---|---|---|---|
| 6th May to 21st May | Community Bonding | Semester exams | 2-3 hrs/day |
| 1st August to 15th August | Work period | College reopens | No effect on time commitment |

The above dates are accurate and there's no way of avoiding them. Apart from that if there are any future clashes with the timeline, it will be communicated timely and it will be made sure that committed working hours to the organization are not disturbed.

**Preferred work hours (CDT): 6.00 am - 2.00 pm (~35-40 hrs/week)**
Along with that, I would be working on weekends also to cover up for any lost time due to time clashes.

# Proposal

**Title: Integrating genomics and high-frequency physiologic data for sepsis detection**

**Overview:**

Sepsis has been defined as organ dysfunction caused by dysregulated immune response to infection [1]. It is one the leading causes of in-hospital deaths among patients worldwide. There are clinical severity scores track a person's status during the stay to gauge the extent of organ failure. There are several machine learning based approaches that consider a patient's clinical information and predicts the occurrence of sepsis in the ICU using labelled training data. Similarly, community approaches [2] to sepsis prediction using gene expression data have developed prognostic models to predict 30-day mortality for sepsis affected patients.

**Work done:**

| Name | Type of data (clinical or genomic) | Description | Prediction task | Model used | Accuracy/Outcome |
|---|---|---|---|---|---|
| Taylor et al | clinical | EHR data for patients visiting the ED | In-hospital mortality | Decision trees, Logistic regression, random forests | Best performing model (random forests) had an AUC of 0.86 |
| Gultepe et al | clinical | EHR data for patients (within 24 hr time bin) who met at least two SIRS criteria | Lactate level<br>In-hospital mortality | Naive Bayes, SVM, GMM, HMM | Lactate prediction using WBC count: AUC: 1.00<br>Mortality: AUC 0.73 |
| ***InSight*** Desautels et al and Calvert et al | clinical | EHR data from MIMIC III restricted to ICU | Sepsis prevalence<br>Sepsis onset | *Insight* | Sepsis Onset AUROC: 0.8799 |
| Mani et al | clinical | EMR for late onset neo-natal sepsis | Sepsis onset | SVM, Naive Bayes, KNN, Random forests | Naive Bayes with AUROC 0.78 which exceeded physician's sensitivity |
| Sutherland et al | Gene expression | Microarray data (65 sepsis) 20 (control) | Biomarkers for sepsis prevalence | Differential gene expression analysis | 42-gene expression markers were identified |

| Name | Type of data (clinical or genomic) | Description | Prediction task | Model used | Accuracy/Outcome |
|---|---|---|---|---|---|
| Scicluna et al | Gene expression | Cohort study for sepsis patients from ICU | Identify biologically relevant molecular endotypes in patients with sepsis. | Unsupervised clustering and classification | Four distinct endotypes were discovered |
| Dickinson et al | Gene expression | Whole blood transcriptome profiling | Identify markers to predict bacterial infection in neonatal sepsis patients | Random forests, SVM, KNN | Identification of a 52-gene classifier that predicts bacterial infection with high accuracy was performed |

**After my initial discussion with Dr. Kamaleswaran, I tried to make a comprehensive map of the type of datasets and the respective metadata for this study. Below are the results of the analysis:**

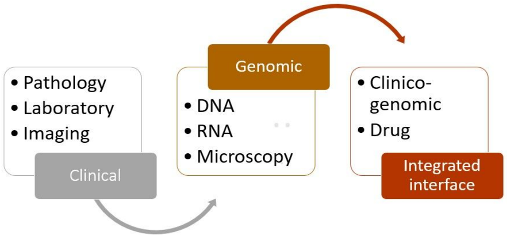
| Dataset | Description | #samples | Phenotypic_data |
|---|---|---|---|
| GSE66890 | Adults in ICU with sepsis +/- ARDS | 57 | Age(real), WBC(real), Sex(categorical), Creatinine(real), ESRD(real), Pneumonia (categorical), Mortality (categorical),apache score(real) |
| GSE66099 | Children in ICU with sepsis/septic shock | 276 | Not available (but this dataset is made from unique samples from 6 other datasets. So requires further investigation) |
| GSE63042 | Adults with sepsis (CAPSOD study) | 129 | Not available |
| GSE40586 | Infants, children, and adults with bacterial meningitis | 40 | Type of bacteria (string) |
| GSE32707 | Adults in MICU with sepsis +/- ARDS | 144 | Not available |
| GSE27131 | Adults with severe H1N1 influenza requiring mechanical ventilation | 21 | Tissue, Gender, Age |
| GSE13015b | NA | 106 | Age, Gender, Ethnicity, Race, Treatment (administered drug info), Pathogen |
| GSE13015a | Adults with sepsis, many from burkholderia | 106 | Age, Gender, Ethnicity, Race, Treatment (administered drug info), Pathogen |

| | | | |
|---|---|---|---|
| GSE10474 | Adults in MICU with sepsis +/- ALI | 34 | Not available |
| E-MTAB-1548 | Adult surgical patients with sepsis (EXPRESS study) | 155 | Age, Gender |
| E-MEXP-3850 | Children w/ meningococcal sepsis | 29 | Not available |
| E-MEXP-3567 | Children with meningococcal sepsis +/- HIV co-infection | 15 | Sex, Developmental stage, Disease, Mortality |
| GSE33341 | Adults with 2+ SIRS criteria and bacteremia | 321 | host strain, host gender, host age, Pathogen, bacterial strain, pathogen dosage, anatomical site, item after infection |
| GSE54514 | Adults in ICU with sepsis | 163 | Age, gender, Tissue, Neutrophil proportion, site of infection |
| GSE63990 | Adults with bacterial infection plus 2+ SIRS criteria | 280 | Not available |
| E-MTAB-4421,E-MTAB-4451 | Adults with sepsis (GAinS study) | 114 | Age, Sex |

**Note:** This table is important in the sense that if we need to know what are all the phenotypic information available with each of these gene expression datasets, it would be an important step in integrating the two. Each of the entries in the Phenotypic data column can be a potential **"missing link"** in bridging the gap between the two forms of data.

**Bridging the gap between genomic and clinical data**

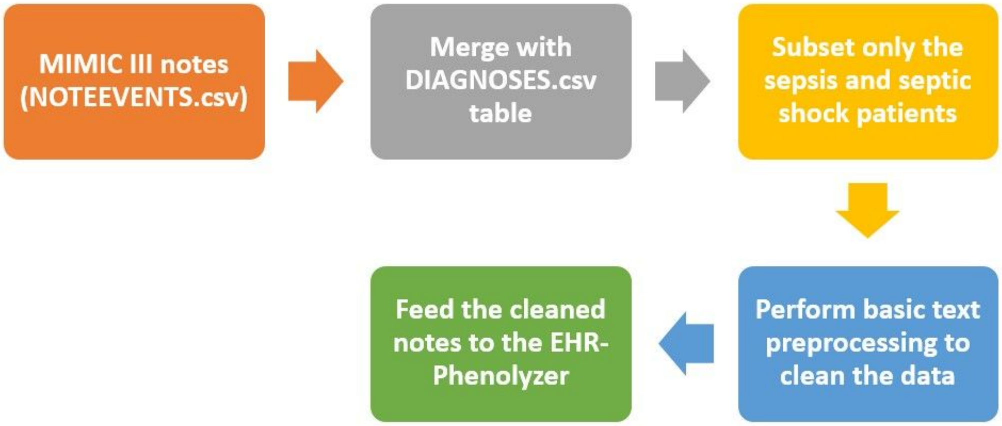
The most important part of this project is to integrate the two data forms (genomic and clinical) using an in-silico approach. There is no dearth of data for each of these individual sources but very hard to find in its combined form . There is a lot of clinical information residing in EHR databases but there is little scope for interoperability between them. Combined with the fact that there are not enough data resources integrating both types of data makes it difficult for clinicians and care providers to make informed decisions. An integrated longitudinal patient record integrating multi-omics will improve disease outcomes drastically. This is one of the major focus areas of this project.



**Tool to combine genomic and clinical data in-silico**

I was trying to build a machine learning model that would train on existing genotypic and phenotypic data and form a map between common disease terms and genes responsible for them. Luckily, I came across a tool that does exactly the same using free text clinical notes. The EHR-Phenolyzer is a python pipeline to automatically translate raw clinical notes into meaningfully ranked candidate causal genes. It might greatly shorten the time for disease causal genes identification and discovery.

Using the MIMIC III clinical notes, I performed a short analysis to extract free text notes for sepsis and septic shock patients only and then run the tool on these notes. The tool outputs two lists: one containing the list of all the phenotypes and the second containing all the causal genes. This is explained in the figure below:


MIMIC III notes (NOTEEVENTS.csv)
Merge with DIAGNOSES.csv table
Subset only the sepsis and septic shock patients
Perform basic text preprocessing to clean the data
Feed the cleaned notes to the EHR-Phenolyzer

The list of genotypic data contains a set of causal genes for every phenotype in the text. So the list of sepsis genes that I got as a result was compared against the top 20 most differentially expressed genes that I would cover later in the proposal. All the genes were present in both the lists. So this shows that the list of causal genes that we got from the tool are indeed differentially expressed in a separate gene expression cohort. This was one significant result that I got from the analysis and perhaps the first step into establishing a link between the clinical and genomic data.

The code used to do the subsetting and preparing the data in the form required for the EHR-Phenolyzer is attached.

The next part of the proposal concentrates on the gene expression data downloaded from the synapse website assigned for the community challenge problem. I was given the task to go through each individual dataset and find a reasonable large sized sample (normal-disease) and find the differentially expressed genes for them.

**Gene expression analysis**

- The principal idea behind doing the gene expression analysis was to confirm the genes that are already mentioned in the literature and to develop a pipeline to analyze the temporal expression profile of these genes, i.e. can we see by taking the SIRS timeframe and compare it to the genes expressed during the Sepsis and Severe Sepsis/Shock phase.
- So the first job was to identify the dataset according to that timeline, i.e. how many samples were collected during SIRS, Sepsis, and Severe Sepsis/Shock. After getting the temporal list of genes in each phase, compare the main genes that are differentially expressed and see which genes have the maximal overlap with those identified in the literature.
- After searching thoroughly through the lists of datasets mentioned in the Synapse website, I came across a longitudinal gene expression dataset GSE13904. Longitudinal analyses were focused on gene expression (of children suffering from sepsis) relative to control samples and patients having paired day 1 and day 3 samples were included in the study. The number of patients found in each phase is given below:
-

| SIRS day1 | Sepsis day1 | Septic shock_day1 | SIRS day3 | Sepsis _day3 | SIRS_resolved _day3 | Septic_shock _day3 |
|---|---|---|---|---|---|---|
| 22 | 32 | 67 | 5 | 20 | 24 | 39 |

**SIRS_resolved was used for patients for not meeting criteria for at least SIRS on day 3**

- So the next job was to perform differential gene expression analysis with the following groups in mind:

| | |
|---|---|
| Control vs Sepsis (Day1) | Control vs Sepsis (Day3) |
| Control vs SIRS (Day1) | Control vs SIRS_resolved (Day3) |
| Control vs Septic shock (Day1) | Control vs Septic shock( Day3) |

**Control vs sepsis (Day1)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| **#upregulated (my analysis)** | **#downregulated (my analysis)** | **#upregulated (Wong et al)** | **#downregulated (Wong et al)** | **#overlap with Wong et al (top 100 upregulated)** ** **and (Names of top 20 genes From my analysis)** | **#overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis)** |
|---|---|---|---|---|---|
| **1066** | **652** | **448** | **158** | **48** | **10** |

**Overlapping (Wong et al and my analysis) upregulated genes: MMP8, CD177, ARG1, MMP9, MCEMP1, ANXA3, IL1R2, OLAH, MMP8 , OLFM4, ANKRD22, VNN1 LTF, IL1R2, OLAH, MS4A4A, GPR84, HPR///HP, HP**
**Overlapping (Wong et al and my analysis) downregulated genes: ZNHIT6 PPP1R16B ANGEL2 PRKCQ FAM162A CDC14A FAM84B NAP1L1 SLC30A4 NUCKS1 NR2C1 ESYT2 IBTK BTBD11 UHRF2 ZMYND11 GORAB DPY19L4**

**Control vs SIRS (Day1)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| #upregulated (my analysis) | #downregulated (my analysis) | #upregulated (Wong et al) | #downregulated (Wong et al) | #overlap with Wong et al (top 100 upregulated) ** and (Names of top 20 genes From my analysis) | #overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis) |
|---|---|---|---|---|---|
| 812 | 412 | 537 | 299 | 59 | 34 |

**Overlapping (Wong et al and my analysis) upregulated genes:**
**MMP8 OLFM4 CD177 ARG1 ANXA3 MCEMP1 MMP9 MS4A4A LTF RETN CRISP3 ANKRD22 HPR///HP GPR84 FCAR HP CLEC5A CLEC4D LCN2**

**Overlapping (Wong et al and my analysis) downregulated genes:**
**FAM162A TN3A3 IL21R TCF3 RHOH LOC100506248 MLLT6 NPM1 MIR6890 //QARS SMYD2 VPS13C RPS24 JMY ZFP62 MRPS18B SNHG1 SNORD83B///SNORD139///RPL3 FAM162A NFATC2 LOC101930583GOLGA8A**

**Control vs Septic shock (Day1)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| #upregulated (my analysis) | #downregulated (my analysis) | #upregulated (Wong et al) | #downregulated (Wong et al) | #overlap with Wong et al (top 100 upregulated) ** and (Names of top 20 genes From my analysis) | #overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis) |
|---|---|---|---|---|---|
| 939 | 852 | 995 | 872 | 42 | 4 |

**Overlapping (Wong et al and my analysis) upregulated genes: MCEMP1 S100A12 CYSTM1 CD177 ANXA3 MMP8 IRAK3 ARG1 S100A9 GYG1 ANKRD22 GPR84 HP DYSF SLC2A14///SLC2A3 HPR///HP SOCS3 IL1R2 CEACAM1 PGD**

**Overlapping (Wong et al and my analysis) downregulated genes: FCER1A KLRF1 MYBL1 MBLAC2 KLHL3 XCL1 YME1L1 SH2D1B YME1L1 CD160 PTPN4 XCL2///XCL1 KLRB1 TRDV3 PTCH1 TC2N KLF12 IKZF2 MGC40069 CD96**

**Day_1_analysis:**

- Number of differentially regulated genes common to all sepsis+shock+sirs (Wong et al): **197**
- Number of differentially regulated genes common to all sepsis+shock+sirs(my analysis): **274**
- Number of overlapping (Wong et al and my analysis) sepsis+sirs+shock genes differentially regulated: **156**
- So, my analysis was able to capture **156** out of **197** reported genes (from Wong et al) that were differentially regulated

**Control vs Sepsis (Day3)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| #upregulated (my analysis) | #downregulated (my analysis) | #upregulated (Wong et al) | #downregulated (Wong et al) | #overlap with Wong et al (top 100 upregulated) ** and (Names of top 20 genes From my analysis) | #overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis) |
|---|---|---|---|---|---|
| 462 | 70 | 150 | 10 | 41 | 8 |

**Overlapping (Wong et al and my analysis) upregulated genes: MMP8 CEACAM8 OLFM4 MMP8 CD177 ARG1 DEFA4 LTF DEFA1B///DEFA3///DEFA1 MMP9ANXA3 VNN1 CRISP3 OLAH MCEMP1 LCN2 IL1R2 MS4A4A LOC149684///BPI ORM2///ORM1**

**Overlapping(Wong et al and my analysis) downregulated genes: FCER1A KLRF1 MYBL1 XCL1 SH2D1B XCL2///XCL1 FCRL3 YME1L1 KLRB1 PDGFD CD160 YME1L1 MAP3K7CL IKZF2 RORA KLF12 EVL S1PR5 TRDV3 TGFBR3**

**Control vs SIRS_resolved (Day3)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| #upregulated (my analysis) | #downregulated (my analysis) | #upregulated (Wong et al) | #downregulated (Wong et al) | #overlap with Wong et al (top 100 upregulated) ** and (Names of top 20 genes From my analysis) | #overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis) |
|---|---|---|---|---|---|
| 675 | 186 | 686 | 213 | 54 | 44 |

**Overlapping (Wong et al and my analysis) upregulated genes: MMP8 OLFM4 ARG1 CEACAM8 CD177 ANXA3 DEFA4 DEFA1B///DEFA3///DEFA1 MMP8 MCEMP1 LTF MMP9 IL1R2 VNN1 BCAT1 MS4A4A**

**Overlapping (Wong et al and my analysis) downregulated genes: MIR4680///PDCD4 CD27 GLS HMGN3 C12orf57 PPM1K PTPN4 ELK4 LDHB ITM2A TSPOAP1-AS1 LCK ETS1 RFTN1 DENND2D FAM69A RUNX3 SGK223**

**Control vs Septic shock(Day3)**
**(Cutoffs: Adjusted p value <0.001 and |log FC| >1)**

| #upregulated (my analysis) | #downregulated (my analysis) | #upregulated (Wong et al) | #downregulated (Wong et al) | #overlap with Wong et al (top 100 upregulated) ** and (Names of top 20 genes From my analysis) | #overlap with Wong et al (top 100 downregulated) and (Names of top 20 genes From my analysis) |
|---|---|---|---|---|---|
| 1124 | 962 | 2150 | 948 | 59 | 58 |

**Overlapping (Wong et al and my analysis) upregulated genes: MMP8 CD177 MMP8 OLFM4 MMP9 ARG1 IL1R2 LTF MCEMP1 OLAH CEACAM8 ANXA3 MS4A4A LCN2 VNN1 HPR///HP IL1R2 RETN HP ANKRD22**

**Overlapping (Wong et al and my analysis) downregulated genes: SRSF8 ANKRD12 FCRL2 TTC19 PRPS1 SNORA16A///SNORA61///SNORA44///SNHG12 CXXC5 ATAD1 CD72 MYC NSMCE4A SHQ1 MAGEH1**

**Day_3_analysis**

- Number of differentially regulated genes common to all sepsis+shock+sirs (Wong et al): **103**
- Number of differentially regulated genes common to all sepsis+shock+sirs (my analysis): **221**
- Number of overlapping (Wong et al and my analysis) sepsis+sirs+shock genes differentially regulated: **81**
- So, my analysis was able to capture **81** out of **103** reported genes (from Wong et al) that were differentially regulated
- Greatest degree of commonality was found in day1 across all three groups: **274**

**Day_1 and Day_3 analysis:**

- Number of genes that are differentially regulated in both day 1 and day 3: **147**

**Work to be done:** The next step would be independently confirm the list of differentially regulated genes' overlap with literature. An extensive data mining operation needs to be performed to mine all the sepsis related DEGs and find the overlap with the **147 common day1 and day3 DEGs** confirmed using our experiment.

Now that we have the list of DEGs from the temporal study, we can use this list to build our machine learning algorithms for the next step of integration.

**Integration of the two datasets: Two possible approaches**

- **Weighted relevant sepsis gene approach**

  This is the most important step of the project. I have come up with a possible approach to deal with this issue. This is an ongoing idea and by no means complete.

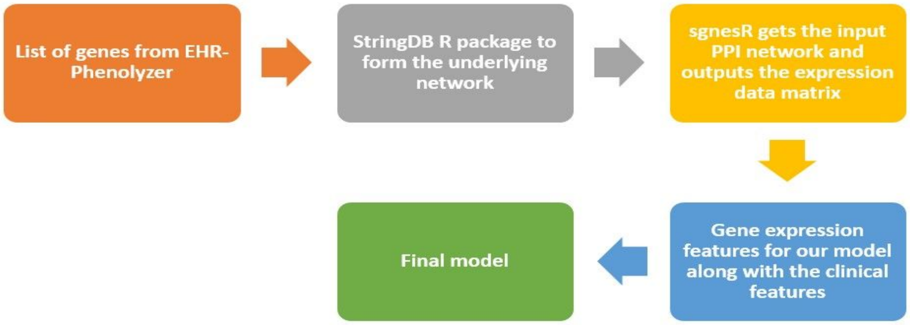
  The idea is to utilize the EHR-phenolyzer to find the list of causal genes and the corresponding clinical parameters like blood lactate level, heart rate, WBC count etc from the MIMIC III database.

| Patient 1 | Free text note 1 | Gene 1 | Gene 2 (sepsis gene) | Gene 3 (sepsis gene) | Gene 4 | WBC count | Blood lactate level |
|---|---|---|---|---|---|---|---|
| Sam | text | 0 | 0.98 | 0.76 | 0 | 9.8 | 0.12 |

**Dummy row for Patient 1 where 2 out of 4 genes were predicted to be causal by EHR-phenolyzer and the rest of the clinical information was mined from his EMR.** The ideas is to rank the predicted sepsis related genes (from EHR-phenolyzer) higher by assigning higher weights to them than the non sepsis related genes. These weights can be learned by the model or assigned prior to training the model. The genes that are not responsible can be given low weights or zero altogether. We can even weigh the genes according to their prevalence in the literature. Predicted genes that are confirmed by 2 or more studies to be involved in sepsis can be given higher weights than the others. Different thresholding criteria can be used to weigh the genes which would now be features to my classifier/regressor. Using say mortality as my target variable, we can derive robust models and see how they perform after the gene information is included.

**Note: This doesn't include the gene expression data in any form. The weights are there to indicate the presence/absence of a sepsis related gene for that patient and how strong is the presence.**

- **Gene expression simulation approach**
  If we have the set of genes from the EHR-phenolyzer, we can take those genes and use the StringDB or the StringDB R-package to construct the underlying PPI network based on selected parameters. After we have the underlying network, we can use sgnesR whose purpose is to enable an easy to use and a quick implementation for generating realistic gene expression data from biologically relevant networks that can be user selected.
  This can be extended to the GSE13904 dataset that we have analyzed earlier. If we get the set of genes from **day_1** and **day_3** separately, we can construct the individual PPI networks and simulate the gene expression data eventually.
  A data object with a network of igraph class can be given as input and a data object which consists expression data matrix can be found as an output. Needless to say, we have to try different benchmarking datasets to test how good our model is performing with respect to other established models out there. If we are not getting any additional increase in our classification accuracy, we might have to think of other methods of including the gene expression features.


List of genes from EHR-Phenolyzer
StringDB R package to form the underlying network
sgnesR gets the input PPI network and outputs the expression data matrix
Gene expression features for our model along with the clinical features
Final model

**Possible hurdles:** The **first** big hurdle is to ensure that our dataset actually mimics the real world. In other words, how likely is that our synthetically generated dataset is similar to actual integrated datasets collected at regular time intervals. We have to make the models generalizable **Secondly**, for time varying datasets it is very important to make sure that the real world minute variances in different time points is actually reflected in our simulated expression data. **Thirdly,** the mapping between the phenotypic and the genotypic data using the tool is strictly based on the presence and absence of certain terms (for patients who already have sepsis) and complex situations like "onset of sepsis" might not be captured. For that we would have to train the algorithm with notes specific to those situations.

## Project plan and timeline

| Days | Work to be done |
|---|---|
| **May 6** | **Student Projects announced** |
| **May 6- May 27** | **Community bonding** |
| May 6- May 21 | ● Introduction with mentors and discussing possible communication medium and time<br>● How to do status updates for the upcoming weeks<br>● Introduction to the project |
| May 22 - May 27 | Reading up relevant research papers on the topic |
| May 28 - June 14<br>(~2.5 weeks)<br>○ ✓ Easy/medium<br>○ ✓ Difficult | ● Revisit the synapse data resource page and do a thorough check(once more) to find longitudinal datasets (SIRS timeline) with enough normal-disease paired samples✓<br>● Collect at least 3-4 such datasets and find DEGs ✓<br>● Report the intersection of genes<br>○ For different timelines ✓<br>○ With literature ✓<br>● Get more sepsis related free text notes apart from the discharge notes MIMIC III database ✓<br>● Revisit the EHR-phenolyzer tool and evaluate the predictions on the new notes ✓<br>● Try to get more notes for patients that were taken before sepsis onset. ✓<br>● Create a customized NLP tool to extract phenotypic terms from natural language of clinical notes that is relevant for different stages of sepsis. ✓<br>● Its very essential to capture the context of the words in the notes to get hold of phrases like "I think I might die" which |

| | |
|---|---|
| | would mean severe pain or discomfort - one of the many sepsis symptoms ✓ |
| June 15 - June 24<br>(~1.5 weeks)<br>○ ✓ Easy/medium<br>○ ✓ Difficult | By this time the preliminary gene expression dataset and the clinical data is expected to be ready to start the analysis.<br>Start with the **weighted relevant sepsis gene approach** and have one set of predictions ready (the target variable can be mortality, sepsis stages etc) ✓ |
| **June 24 - June 28** | **First evaluations** |
| June 29 - July 13<br>(2 weeks)<br>○ ✓ Easy/medium<br>○ ✓ Difficult | ● Based on the previous integration approach, we would move on with the second part, i.e **simulation of gene expression** ✓<br>● First job would be to make sure that the distribution of the simulated gene expression values should be similar (mean and variance) to what we usually observe with time series gene expression data for sepsis patients ✓<br>● Form PPI networks using the list of genes as input (stringDB R-package) ✓<br>● From the PPI networks, feed the igraph object to the sgnesR package and tune the parameters to take care of the intended variation (as explained in point 2 above) ✓<br>● Test our two models' accuracy with literature and see how they are performing as compared to the models based on clinical severity scores (like SOFA etc) ✓ |
| July 14 - July 21<br>(1 week) | Buffer week |
| **July 22 - July 26** | **Second evaluations** |

| | |
|---|---|
| July 26 - August 9<br>(2 weeks)<br>● ✓ **Easy/medium**<br>● ✓ **Difficult** | ● Disease subtyping is an essential component to understanding complex diseases. With the integrated data at hand, I would try to do some unsupervised clustering techniques to see, if the data separated out into novel clusters and subgroups. Verification of older subtypes can also happen. Newer subtypes might also emerge as a result.✓<br>● Finding the corresponding clinical parameters for these novel clusters will be the next step. ✓<br>● This experiment can be repeated for both adult and neonatal sepsis categories. ✓ |
| August 10 - August 18<br>(1 week) | ● Code cleaning and documentation and writing up reports mentioning the results |
| **August 19 - August 26** | **Students submit code and evaluations** |
| **August 16 - September 2** | **Mentors submit final evaluation** |
| **September 3** | **Results announced** |

**Note:**

- The above timeline is a tentative one and I am open to change based on the mentor's feedback
- Holidays are also included while writing the above timeline
- GSOC timeline events are marked in grey and merged with mine for better interpretability
- Buffer period have been kept to cope up with any unforeseen events

## Conclusion and note of thanks

I am truly delighted to have been given the opportunity to write the proposal for CBMI @UTHSC for this year's GSOC. I sincerely appreciate Dr. Mohammed and Dr. Kamaleswaran for their constructive feedback during the proposal writing and ideation  phase.
I am really looking forward to any feedback from the organisation members reviewing this document, and would be glad to discuss/change accordingly.
Finally, it would be very exciting to work on this challenging project this summer, given the opportunity.