Data integration and visualization for complex placenta protein data

Prof. Dr. Bernhard Renard, Elizabeth Yuu
Data Analytics and Computational Statistics

How to impact

Data visualization and **integration** are essential in the field of research. It is imperative to find ways to efficiently share and present data. By doing so, we enable other researchers to build upon and continue previous works and pursue novel ideas.

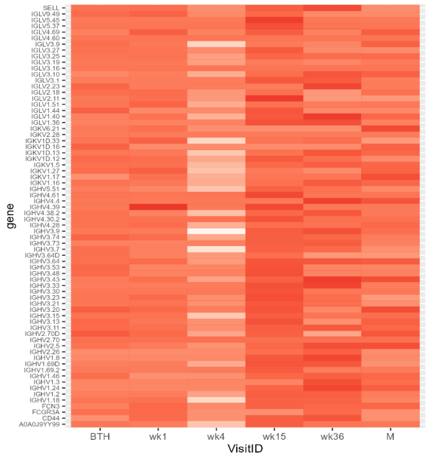
Dashboards are extensively utilized in research to provide easily accessible visual representations of data. They enable users to interact with dynamic information from multiple sources in a single location. Additionally, dashboards aid in facilitating **statistical analyses** of the acquired data. For this project, you will develop a dashboard to present placenta protein data.

Clinical data

Placenta protein data from women and infants in Africa

Have you ever wanted to work with real world clinical data or learn how to effectively visualize a data's story? Then this is your project.

Studying human placenta proteins is not a novel field of research, yet it is not thoroughly explored. We are eager to share with you our dataset that is comprised of blood samples from pregnant women and their infants from Africa. Some of the mothers were HIV positive and received medical treatment in the clinical study. In turn, some infants were also exposed to the virus and the treatment.



Heat map of differing concentrations of proteins at various time points: at birth, week 1, week 4, week 15, week 36, and mother base line. Potential research question: why did week 4 have the least concentration of proteins but had stronger concentrations at later time points?

Project tasks

Visualization & Statistics

Project goal: Create intuitive and interactive visuals of complex data that will serve as a resource on placenta protein data with the option of performing statistical analyses.

We envision having two groups for this project, one group for the dashboard and one group for the statistics aspect.

Together, the two groups will design a platform for future researchers to reference when performing placenta protein related research.

Dashboard:

- **Web scrape** information related to placenta proteins. This includes research papers, articles similar to our clinical data and placenta protein specific information.
- **Optimize** web scraping process from multiple resources, for example from Pubmed, UniProt and Google Scholar.
- Adjust the frequency of when to scrape particular information, such as scraping for papers on a weekly basis vs scraping for protein information on a monthly basis.
- **Create and design a user-friendly interactive interface** for the data that is accessible to broad audiences. In other words, how to best visualize the complex data in a form that is easily understood.

- **Statistics:**
- **Hierarchal clustering** -> identify which proteins cluster together and why?
- **Data visualization** -> effectively visualize specific statistical conclusions ie. varying correlation trends between time points, missing data, and comparative statistics between datasets.
- **Summary statistics** -> demonstrate significant findings, specifically on the clusters, correlations and trends. You will not only report these statistics, but you will learn which statistics are best to report when working specifically with protein data.

Your skill set

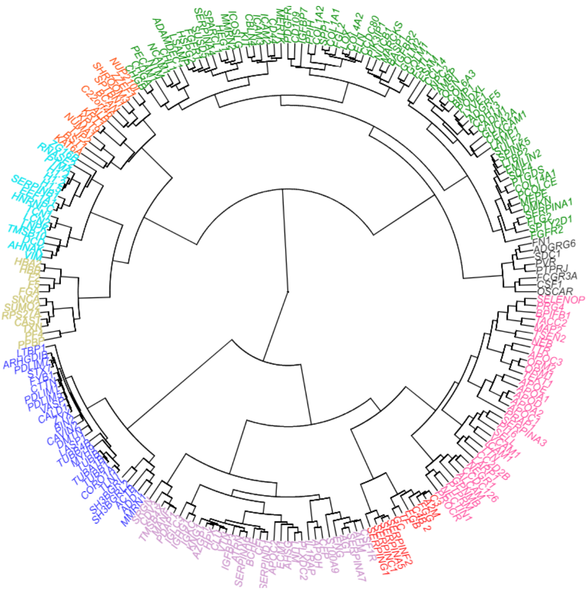
Mandatory requirements:

- Strong Python / R programming
- Familiarity with Dash or R Shiny or similar dashboard generating related tool
- Minimum have taken 1 statistics related course or has experience in statistics/data analysis
- Knowledge of biology is not required, but a curiosity to learn about biology, is desirable

Information

- Bernhard.Renard@hpi.de
- Elizabeth.yuu@hpi.de

Maximum 4 students



Hierarchal clustering of the proteins. You will have the opportunity to further explore all or individual clusterings. For example, why do these proteins cluster, what do they have in common, or do all clusterings follow the same trends?

Why this project?

- You will work on a heavily **applied project** with **real clinical data**
- You will have **first hand experiences** working with **missing**, **complex data** and learn how to share your results with other researchers.
- You will have opportunities to explore **your own questions** and **ideas** that may manifest during the project timeline. Of course we will provide as much **support** and **guidance** as needed to **support** your creativity in this project.

Additional sources:

NSF-powered dashboards you should bookmark: **https**://new.nsf.gov/science-matters/nsf-powered-dashboards-you-should-bookmark

R Shiny: https://shiny.posit.co/

Dash: https://dash.plotly.com/

Placenta-derived proteins across gestation in healthy pregnancies—a novel approach to assess placental function?:https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-022-02415-z