Human Genetics Final Project Statistical Analysis Plan:

Aidan Neher - November 17, 2023

**Abstract:** Effective variable selection and prediction methods, exemplified by recent advancements like Grace-AKO (1) and BIPnet (2), are crucial in identifying disease-associated genes for clinical outcomes. This study primarily aims to compare Grace-AKO and BIPnet in variable selection and assess their predictive abilities. Using simulated data and potentially breast cancer gene screening data, we’ll present expected results summarizing variable selection and clinical prediction in tables. Challenges may arise from differing treatments of network information and multiple data types in BIPnet, as well as access limitations to clinical data with specific characteristics.

**Background:** Variable selection and prediction are crucial in identifying disease-associated genes for clinical outcomes. Grace-AKO and BIPnet leverage prior network information for this purpose. Grace-AKO integrates multiple knockoffs within a network-constrained penalty to control FDR in high-dimensional settings. In contrast, BIPnet performs Bayesian integrative analysis and prediction using multiple omics data and allows for clinical covariates. This study aims to compare Grace-AKO and BIPnet in variable selection performance and explore their predictive abilities.

**Research Question:** How do Grace-AKO or BIPnet differ in variable selection and prediction?

**Data:** Firstly, simulated data will be analyzed. We will simulate two datatypes . Each have n rows representing an observation and p features. . For each datatype, we let the first 100 variables form the networks, 10 of which are main variables, each connected to 9 variables. Each datatype’s resulting network has 100 variables and edges, and p − 100 singletons. The network structure is captured by the covariance matrices:

is block diagonal with 10 blocks of size 10, between-block correlation 0 and within each block, the correlation between a main and each of its 9 connecting variables is 0.7, and the correlation between each of the 9 connecting variables is represented by a 9×9 compound symmetric submatrix with correlation 0.49. For BIPnet analysis, these two datatypes will be kept separate and used to simulate the latent factor, factor loadings, and outcome variable. For Grace-AKO analysis, since it is specific to use in one datatype, we will concatenate the to form one omics matrix that can be used for outcome simulation.

Time-permitting, we will assess these methods using clinical data. Specifically, we will use the readily available breast cancer gene screening data, which includes gene expression microarray data for 645 genes, from 348 breast cancer tumors from different individuals (3). For each gene (row) and sample (column) the given value is a measure of how active that gene is (“expression”) for that sample. The clinical subtype for each sample is given as “Basal” or “non-Basal”. Basal tumors have a poorer prognosis and respond to certain therapies differently.

Gene network information is more difficult to source. The Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) is seemingly behind a paywall and several of the R packages traditionally used to access it are now deprecated. Furthermore, the University of Minnesota no longer holds a license to access Ingenuity Pathway Analysis (IPA), a software program which can analyze the gene expression patterns using a built-in scientific literature based database (IPA, www.ingenuity.com). The ability to access this information will also determine whether this project will include clinical data analysis.

**Primary Outcomes:** FDR, power, and number of variables selected will be used to assess the differences in variable selection performance. **Secondary Outcomes:** MSE will be used for simulated continuous outcome. AUC will be used for each sample’s clinical subtype prediction.

**Statistical Methods & Software:** Grace-AKO and BIPnet will be used for variable selection in the simulated data, and the primary outcomes of interest will be evaluated. In the context of clinical data, first Grace-AKO and BIPnet will be used for variable selection and then the selected features will be used for logistic regression prediction of Basal or non-Basal clinical subtypes. All analyses will be conducted using the R statistical software, version 4.2.0.

**Table 1: Simulated Data Variable Selection Table Shell:**

| Model | FDR | Power | N Variables Selected |
| --- | --- | --- | --- |
| Grace-AKO |  |  |  |
| BIPnet |  |  |  |
| Difference |  |  |  |

**Table 2: Clinical Prediction Results Table Shell:**

| Model | AUC-ROC |
| --- | --- |
| Grace-AKO | (95% CI) |
| BIPnet |  |
| Difference |  |

**Potential Challenges:** Incorporating prior network information poses challenges for both simulated and clinical data analyses. Grace-AKO accommodates edge weights in networks, unlike BIPnet, but setting uniform Grace-AKO network edge weights may ensure comparability. Additionally, Grace-AKO’s focus on a single datatype complicates creating universally applicable simulated data. Obtaining clinical data with necessary attributes—like prior grouping information, multiple omics types (for BIPnet), and continuous outcomes—poses challenges. The available breast cancer data only offers binary outcomes, necessitating a two-step variable selection and logistic regression analysis.

**References:**

1. Tian, P., Hu, Y., Liu, Z. & Zhang, Y. D. Grace-AKO: a novel and stable knockoff filter for variable selection incorporating gene network structures. BMC Bioinformatics 23, 1–16 (2022).
2. Chekouo, T. & Safo, S. E. Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. Biostatistics 24, 124–139 (2021).
3. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).