Approaches to Incorporating Whole Genome and Multi-Omics Information in the context of Clincal Prediction

Siddharth Avadhanam
avadhana@msu.edu

1. PROBLEM DESCRIPTION

The high-dimensionality and complexity of data coming in from Genome Wide Association Studies poses a whole range of statistical and computational challenges in finding genetic covariates of important phenotypes (disease, yield , height etc). These data commonly contain thousands of individuals with information on hundreds of thousands of markers. [1] An important issue which has been given considerable attention in the statistical genetics literature is how to choose a subset of these markers so that the analysis of these data with standard commodity hardware and statistical methods becomes more tractable.[2] Methods which analyze a single-marker at a time are popular, but are severly limited in that they do not take into account interactions between markers, and come with their own set of challenges related to multiple testing.[3] It is thus becoming increasingly important to explore methods that incorporate information form the whole genome and possibly from outside the domain of traditional statistics.

Further, while methods that incorporate information form the whole genome simultaneously are finding increasing success, there is only so much that can be accomplished by solely considering genetic marker information. These approaches as a whole are running into the problems of missing heritability and are able to explain only a small proportion of the inter-individual variation for multifactorial outcomes such as height.[4] The complexity of the underlying biology demands approaches that can synthesize data from multiple platforms and consider factors contributing to disease that lie outside the genome. This more biologically realistic approach would include data sources as diverse as epigenetics, pedigree data, transcriptomic data, proteomic data etc.[5] Here we will consider incorporating epigenetic and pedigree data in addition to genomic data.

Finally, we will be considering Kernel based approaches within and outside of the bayesian paradigm and how they can be applied to genomic enabled clinical prediction. Kernel methods have immense flexibility (can be adopted towards multiple data-types) and are particularly useful for high-dimensional problems.[6]

In this work I propose to evaluate bayesian methodology and contrast it with classical approaches in the context of multi-omic clinical prediction. The objective here will be two fold. Firstly, We will evaluate methods to incorporate information from the entire genome simulataneously (the so called whole-genome approach) as opposed to one marker at a time. Second, we will look at methods for bringing in multiple-layers of information (multiomics prediction)

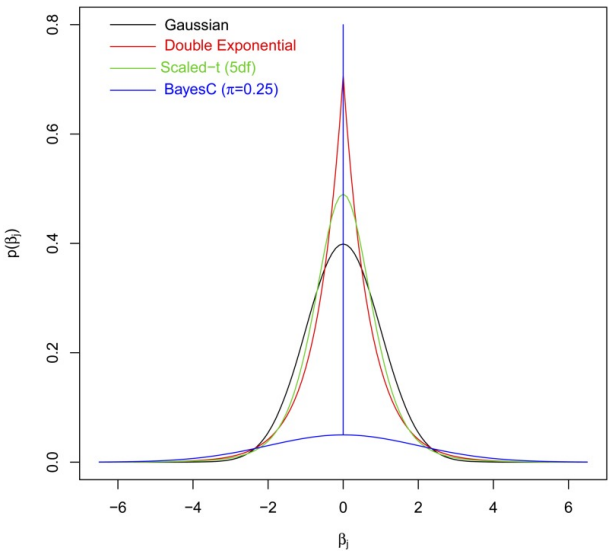

Figure 1: Figure 1 by Perez et al.[8] shows different prior densities implemented in BGLR

which will include genetic markers, pedigree networks, and epigenetic data. We will see how these methods explain a higher proportion of the phenotypic variance, as well as give us more accurate estimates of patient survival. Classical whole genome regression approaches we might look at include methods such as the lasso and ridge-regression.[7]

All methods will be implemented in R. Bayesian methods will be executed using the BGLR (Bayesian Generalized Linear Regression) package by Perez et al.[8] which provides a powerful modelling framework for incorporating bayesian regressions. Differential shrinkage and marker selection can be implemented with an adequate choice of prior provided as options in the package, as well as the ability to include single or multi-kernel methods.

A good example of the motivation behind this work can be found in Vazquez et al. [9] I will implement some of the more succesful approaches in a dataset obtained from specific cohorts of the Framingham Heart Study, which includes phenotypic data, demographic data, genotyping data, pedigree information, and DNA-methylation data.

2. RELATED WORK

There is considerable literature in statistical learning approaches towards the analysis of GWAS data. Earlier work

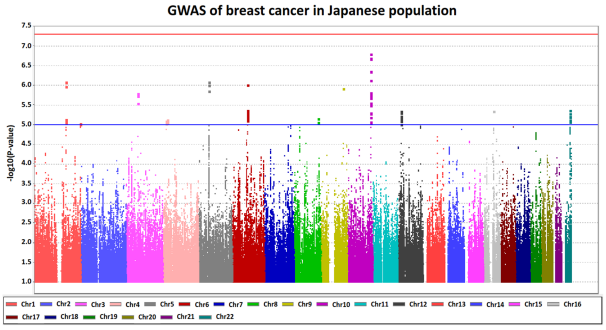

Figure 2: Figure 2 shows an example manhattan plot by Low et al.[10] constructed from single marker regression across the genome. The y-axis is the $-log(pvalue)$ and the x-axis contains SNP positions colored by chromosome.

has touched upon single marker regression methods and the whole host of approaches that can be applied to deal with the multiple-testing issues.

Penalized regression methods such as ridge regression,[11] variable selection methods such as the lasso and dimensionality reducing approaches such as principal components [12] have all been applied and studied extensively in the analysis of genome-wide data.

The discovery of gene-gene interactions is an active area of research where machine learning tools are being extensively employed [13] Bayesian methods and mixed models approaches within specific communities such as animal breeding have been studied and applied for a long time, and are increasingly being used with genome-wide data. [2].

The issue of missing heritability and multi-factorial traits have gotten considerable attention in the literature. Multi-omics approaches toward improving prediction accuracy are starting to gain traction with increase in the availability of computational resources. In addition to better prediction, these approaches can be particularly illuminating when it comes to the biological mechanisms being studied. [5] An example would be to try and understand how much of the variation in the trait of interest can be explained by epigenetic variation, and is one of the things we are interested in here.

Another direction in the literature makes use of the methods mentioned so far to better understand the pathogenetic etiology of disease by studying genetic overlap between the disease and it's comorbidites.

3. DATA DESCRIPTION

This project will require the use of heterogenous data types. The Genotyping information is stored as binary files, which needs to be accessed by a custom script which reads information from metadata associated with the file. The large amount of data collected from a study of this size are organized by filename in a study manifest. Finding and collecting together the requisite data is itself a challenge and needs to be done systematically.

Phenotypes are stored and organized by study cohort, examination number and consent group. Variable checks need to be done before such data is extracted to ensure the right variables are present. Demographic information is stored seperately and are also organized the same way. We also incorporate Pedigree information into the analysis, which can be though of as a directed graph representing strengths of relationships between individuals.

Finally, methylation data needs to be extracted and incorporated into the analysis. I am still at the stage of validating this information, and will report further progress in the final report.

4. PROJECT MILESTONES

This section will detail milestones so that I can roughly organize my work toward the completion of this project. I am at the post Data-Extraction and Cleaning stage. This step took some time as there are multiple different data types (Genotypes, Phenotypes, Demographic, Methylation and Pedigree). These include different consent groups, binary files, metadata, and data manifests.

4.1 Completed Milestones

- Prepare a proposal and survey important literature.
- Dig deeper into the literature and search for implementational details.
- Identify relevant software. R package BGLR will be used extensively to implement computationally intensive Bayesian methods.
- Extract the data based on file-name lookup on study manifest file.
- Clean, organize and explore the data. Include or remove clinical covariates as necessary.
- Obtain metadata required for reading Binary files. This required searching the Affymetrix website for the corresponding genotyping array and the marker information.
- Read in the Binary Genotype files and prepare the relationship matrices. This step was computationally intensive and the MSU High Performance Computing Cluster was used to prepare the relationship matrices. The Relationship matrix was computed as

$$R = XX^T \tag{1}$$

  Where X is a high dimensional $9000 * 50000$ matrix containing marker information. More details on the role of the relationship matrix in inference will follow in the final project report.
- Prepare a second draft with a detailed introduction and some results based on exploration.

4.2 Remaining Milestones

- Prepare the code, run and test it extensively on example data. Decide if simulations will be necessary.
- Run the alogrithms on the Framingham data set, and collect the results. Prepare summaries and visualizations.
- Prepare Final Report

5. REFERENCES

[1] Xiang Zhang, Shunping Huang, Zhaojun Zhang, and Wei Wang. Chapter 10: Mining genome-wide genetic markers. 8(12):e1002828.

[2] Gustavo de los Campos, John M. Hickey, Ricardo Pong-Wong, Hans D. Daetwyler, and Mario PL Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. 193(2):327–345.

[3] Valentina Moskvina and Karl Michael Schmidt. On multiple-testing correction in genome-wide association studies. 32(6):567–573.

[4] Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. 11(6):446–450.

[5] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. 7:13090.

[6] Xuefeng Wang, Eric P. Xing, and Daniel J. Schaid. Kernel methods for large-scale genomic data analysis. 16(2):183–192.

[7] Patrik Waldmann, GÃąbor MÃľszÃąros, Birgit Gredler, Christian Fuerst, and Johann SÃűlkner. Evaluation of the lasso and the elastic net in genome-wide association studies. 4.

[8] Paulino PÃľrez and Gustavo de los Campos. Genome-wide regression and prediction with the BGLR statistical package. 198(2):483–495.

[9] Ana I. Vazquez, Yogasudha Veturi, Michael Behring, Sadeep Shrestha, Matias Kirst, Marcio F. R. Resende, and Gustavo de los Campos. Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. 203(3):1425–1438.

[10] Siew-Kee Low, Atsushi Takahashi, Kyota Ashikawa, Johji Inazawa, Yoshio Miki, Michiaki Kubo, Yusuke Nakamura, and Toyomasa Katagiri. Genome-wide association study of breast cancer in the japanese population. 8(10):e76463.

[11] Erin Austin, Wei Pan, and Xiaotong Shen. Penalized regression and risk prediction in genome-wide association studies. 6(4).

[12] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. 38(8):904–909.

[13] Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege, and Andrew Collins. Machine learning approaches for the discovery of geneâĂŞgene interactions in disease data. 14(2):251–260.