Neural Network-Based Multiomics Data Integration in Alzheimer's Disease

Pankhuri Singhal*, Shefali S. Verma, Scott M. Dudek, Marylyn D. Ritchie

Perelman School of Medicine, University of Pennsylvania

*Corresponding Author: singhalp@pennmedicine.upenn.edu

ABSTRACT

Alzheimer's Disease (AD) is a growing pandemic affecting over 50 million individuals worldwide. While individual molecular traits have been found to be associated with AD at the DNA, RNA, protein, and epigenetic level, the underlying genetic etiology of AD remains unknown. Integrating multiple omics datatypes simultaneously has the potential to reveal interactions within and between these molecular features. In order to identify disease driving mechanism, a standardized framework for integrating multiomics data is needed. Due to high variability in size, structure, and availability of high-throughput omics data, there is currently no gold standard for combining different data types together in a biologically meaningful way. Thus, we propose a pathway-centric, neural network-based framework to integrate multiomics AD data. In this knowledge-driven approach, we evaluate different gene ontologies to map data to the pathway level. Preliminary results show integrating multiple datatypes under this framework produces more robust AD pathway models compared to models from single data types alone.

KEYWORDS

Multiomics, Data Integration, Curse of Dimensionality, Neural Networks, Biological Application

ACM Reference format:

P. Singhal, S. Verma, S. Dudek, and M. Ritchie. 2019. Neural Network-Based Multiomics Data Integration in Alzheimer's Disease*, Prague, Czech Republic, July 2019 (GECCO'19)*, 2 pages. https://doi.org/10.1145/3319619.3321920

1 INTRODUCTION

Rapidly declining costs of high-throughput instruments has yielded a proliferation of experimentally derived multiomics data. High dimensional multiomics data of different sizes and sparsity have become available and multiomics data of different sizes and sparsity have become available and are being used to investigate genetic architecture underlying complex traits. However, most methods use a step-wise integration strategy, evaluating a single omics datatype a time, and thus fail to account for complex interactions occurring between different levels of gene regulation[1]. Thus, systematic integration of individual-level data is needed to gain insight into biological mechanism and reveal complex predictive patterns in clinical outcomes. Existing methods allow homogenous data, or omics datasets of the same type, to be integrated across studies. The lack of standardized methods for integrating multiomics data, both across and within studies[2], has become more pronounced now as data-mining for omics imputation becomes increasingly accepted in genomics and bioinformatics communities. Given this, there is a need for a normalization framework that can integrate heterogenous data without compromising accuracy and without increasing type 1 error.

1.1 Alzheimer's Disease

To date, genome-wide association studies have identified over 500 candidate genes in AD. However, the few genes that have replicated explain disease in a fraction of the AD population. Evaluating single nucleotide polymorphism (SNP) variation alone has not given us the full picture of AD mechanism. Furthermore, while the hallmark amyloid beta plaque pathology in AD has been well-characterized, there are different molecular pathways across individuals posited to play a role in driving this common pathology[3]. Thus, to identify causal variation beyond the SNP level and to interrogate pathway heterogeneity underlying AD, integration of multiomics AD data is needed. In this study, we propose using a Grammatical Evolution Neural Network (GENN) to integrate multiomics data from the Religious Order Study and Memory and Aging Project (ROSMAP). Previous studies[6] have shown GENN outperforms other machine learning methods in generating accurate interaction models, given robust approximation of correlation between features.

1.2 The "Curse of Dimensionality"

While the explosion of omics data hides more clues for understanding the underlying mechanism of complex traits, it also presents an analytical challenge in building meaningful statistical models to identify true significant variables. This "Curse of Dimensionality" hinders evaluation of all possible combinations of molecular features given computational limitations with respect to combinatorics and data sparsity. Previous studies have shown that reducing dimensionality of quantitative omics data by aggregating values, or taking the mean, across pathways to

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3321920

generate "pathway scores" results in more accurate models of binary clinical outcome compared to full omics datasets alone[4]. Thus, we propose 1) using a variety of gene ontology knowledgebases to generate pathway scores for methylation, gene expression, and protein expression data and 2) simultaneously integrating all pathway scores using GENN.

2 METHODS

From the ROSMAP dataset, we created a case-control cohort of 435 individuals who all had methyl array and RNA-seq data. Individuals with a cognitive score of 1-3 (mild cognitive decline and non-AD dementia) were classified as controls and individuals with a cognitive score of 4 or 5 (advanced cognitive decline and AD-dementia) were classified as cases. We first mapped probes from both data types to gene annotations; intergenic probes were mapped to the closest gene within 50kb. We then used biological knowledgebases Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Reactome to separately map genes to pathways. The KEGG annotated data mapped to 266 pathways, GO mapped to 315 pathways, and Reactome mapped to 247 pathways. To generate pathway scores, we first log transformed the data to normalize it, and then obtained the mean of the genes in each pathway. These means functioned as pathway scores, representing variation in aggregate for a single pathway in a datatype.

2.1 Grammatical Evolution Neural Network

We tested the hypotheses 1) models from integrated methylation and gene expression data have higher accuracies compared to models from each data type alone and 2) gene-pathway annotations from different knowledgebases contribute a bias in models by affecting error or accuracy. To test these hypotheses, we used each of the three knowledgebases separately to combine pathway scores from each data type alone and then combined. In total, we test 9 different models. Specifically, we used GENN machine learning tool called Analysis Tool for Heritable and Environmental Network Associations[5] (ATHENA) to integrate pathway scores. We assessed balanced accuracy, type 1 and 2 error, and area under the curve (AUC) to evaluate models. The details of the grammatical evolution algorithm for GENN are: 1) The dataset is divided into five equal parts for 5-fold cross-validation (4/5 for training and 1/5 for testing). Training begins by generating a random population of binary strings initialized to be functional GENNs. Both the model structure and the variables included are randomly generated. 2) The GENNs in the population are evaluated using the training data and the fitness for each model is recorded. The solutions with the highest fitness are selected for crossover and reproduction, and a new population is generated. 3) Step 2 is repeated for a predefined number of generations. The overall best solution across generations is tested using the remaining 1/5 data and fitness is recorded. 4) Steps 2–4 are repeated four more times, each time using a different 4/5 of the data for training and 1/5 for testing. The best model is defined as the model identified the most over all five cross-validations. Network parameters were optimized.

3 EXPERIMENTAL RESULTS

ROSMAP Data type	Biological Knowledgebase	Balanced Accuracy	Pathways in Model
Methylation Array	GO	0.586	3
Methylation Array	KEGG	0.512	7
Methylation Array	REACTOME	O.479	6
RNA-seq	GO	0.598	8
RNA-seq	KEGG	0.612	5
RNA-seq	REACTOME	0.620	5
Methylation Array and RNA-seq	GO	0.676	11
Methylation Array and RNA-seq	KEGG	0.591	6
Methylation Array and RNA-seq	REACTOME	0.688	8

Table 1: GENN models for single vs. integrated datasets and corresponding knowledgebases

4 CONCLUSIONS

Preliminary results from this study indicate that using a knowledge-driven approach to integrating methylation array data and RNA-seq data via GENN results in more accurate models compared to models using only a single data type. Table 1 depicts the average accuracy of the integrated models to be 0.652, while the average accuracy of the methylation data models is 0.526 and the expression data models is 0.61. These early findings indicate that the differences in pathway annotations across these databases can affect the accuracy of the pathway interaction models. The integrated models from both GO and KEGG contained the "Vasopressin-regulation" pathway that has not previously been cited to play a role in AD to the best of our knowledge. Additionally, all three integrated models were enriched with putative AD pathways from the literature such as "Parkinson's Disease" pathway, "Cytokine-signaling" pathway, and "Complement cascade" pathway. These positive results function as a validation and suggest our proposed method has the power to identify complex interactions between data types that have been missed due to a single data type approach. We will conduct further study with larger multiomics datasets and benchmark our approach against other integration methods utilizing pathway aggregation.

REFERENCES

[1] Das, S., Majumder, P. P., Chatterjee, R., Chatterjee, A., & Mukhopadhyay, I. (2018). A powerful method to integrate genotype and gene expression data for dissecting the genetic architecture of a disease. *Genomics*, (March), 1–8.

[2] Kang, M., Kim, D., Gao, J., et al. (2017). Integration of multi-omics data for integrative gene regulatory network inference. *International Journal of Data Mining and Bioinformatics*, *18*(3), 223.

[3] Iqbal, K., Liu, F., Gong, C. X., & Grundke-Iqbal, I. 2010. Tau in Alzheimer disease and related tauopathies. *Current Alzheimer research*, *7*(8), 656-64.

[4] Kim, D., Li, R., Dudek, S. M., Frase, et al. (2014). Knowledge-driven genomic interactions: An application in ovarian cancer. *BioData Mining*, *7*(1), 1–11. Holzinger, E. R., Dudek, S. M., Frase, A. T., Pendergrass, S. A., & Ritchie, M.

[5] D. (2014). ATHENA: The analysis tool for heritable and environmental network associations. *Bioinformatics*, *30*(5), 698–705.

[6] Kim, D., Li, R., Lucas, A., Verma, et al. (2017). Using knowledge-driven genomic interactions for multi-omics data analysis. *Journal of the American Medical Informatics Association*, *24*(3), 577–587.