Human gut microbiome for pediatric Autism Spectrum Disorder (ASD) Report

Gutsy Graduates: Yixiang Qu, Jialiu Xie, Jose S. Lopez, Yifan Dai, Jane She

2023-04-09

Introduction

Backgound & Motivation

Autism spectrum disorder (ASD) is characterized by behaviors in the areas of social communication and restricted repetitive sensory-motor behaviors. While ASD results from altered brain development and neural reorganization, there are no genetic biomarkers for diagnosis. Instead, diagnosis is characterized by a patient's behavior which are assessed on different scales by physicians through clinical observation as well as caregiver reports. However, diagnosis screening processes are often different for young children, older children, adolescents, and adults. The majority of children with ASD in North America and northern Europe will have already been diagnosed by early school age, but there are still those without a diagnosis (Lord et al. 2018). Validity of self-report and self-diagnosing instruments are also questionable.

In 2012, it was estimated by the WHO that the global prevalence of ASD was around 1% which was concluded based on reviewing epidemiological surveys worldwide (Elsabbagh et al. 2012). While it was once considered rare, current estimates suggest that these rates have risen in prevalence, and there is controversy as to what is causing these increased rates (Matson and Kozlowski 2011). These changes may possibly be attributed to new assessment instruments, improved awareness, and changes in diagnostic criteria.

Despite the increased awareness about the disorder, its underlying causes remain unclear. There have been genetic components linked to ASD, but determining the specific mutations associated with ASD is very difficult due to the high variation of underlying gene mutations (Sutcliffe 2008). Additionally, such genetic anomalies only account for a small fraction of cases, around 7% from current estimates (Landrigan 2010). However, this small percentage is also due to currently undiscovered genetic associations and is expected to increase with the advancement of genetic research. It has also been hypothesized that environmental exposures may also contribute to the causation of the disease.

One of the commonly seen comorbidities in patients with ASD is the occurrence of gastrointestinal (GI) related disturbances. Recently, links have been formed between the human gut microbiome and ASD

symptoms (Vuong and Hsiao 2017). Changes in the gut microbiome have been demonstrated to modulate behavior through the gut-microbiome-brain axis.

Microbiome data is known to be challenging to work with for several reasons. Microbiome data faces challenges in data quality, as different measurement techniques from experimental replicates may yield high variability in results. Additionally, data is high dimensional naturally inner-correlated, violating the independence assumption in many traditional analysis approaches. Zero-inflation is also present in the data due to zeroes from sampling, rounding, and normalization from preprocessing. Finally, microbiome data is always positive, so users need to be intentional in choosing a modeling strategy that allows for positive results.

To tackle the issue of high dimensionality, dimensionality reduction techniques have been applied using methods such as principal component analysis (PCA). More recently, preprocessing methods for microbiome datasets have been developed which borrow ideas from the field of natural language processing (NLP). (Tataru and David 2020) utilized the GloVE word embedding algorithm to preprocess microbiome data prior to using a random forest based method to predict a binary outcome. This paper concluded that performanced based on NLP-inspired preprocessing was superior to that of PCA.

Data Overview

The dataset was found on Kaggle, which originally drew from (Dan et al. 2020), and contains information from a cohort of 143 Autism Spectrum Disorder (ASD) patients and typically developed (TD) patients spanning ages from 2-13 years old which were recruited from May, 2016 to August, 2017. The 143 ASD patients were age and sex-matched with the 143 TD children who were recruited. Additionally, metagenomic analysis of gut microbiota was performed for 30 constipated ASD patients and 30 TD patients, as constipation is another major symptom presenting in ASD patients.

Kaggle has available a 16S rRNA sequence of gut microbiota with 1322 rows and 256 columns. The rows are numbered according to Operational Taxonomic Units (OTUs) which are used to classify groups of closely related individuals, similar to the concept of Linnaean taxonomy or evolutionary taxonomy. OTUs construct "mathematically" defined taxa and their use is widely accepted and applied to describe bacterial communities using amplicon sequencing of 16S rRNA gene (Lladó Fernández, Větrovskỳ, and Baldrian 2019). So, there were 1322 of these OTUs identified. The columns then include a "taxonomy" column, which are lists of the microbiotic species belonging to that OTU. For the rest of the columns, we have 143 columns denoting sample IDs from TD children, and 111 columns denoting sample IDs from ASD children. Down the columns as observations, we see numbers corresponding to each row (OTU) which indicate a measurement of gene abundance.

Additionally, there is a second dataset which contains data from the metagenomic analysis performed for 30 ASD patients and 30 TD patients. The data is presented similarly with patient IDs for 30 ASD and 30 TD children across the top. There are 5619 rows which identify different bacteria present, this time as individual rows rather than in groups like before. Similar to the previous dataset, down the columns as observations, we see measures of abundance in patients. For our project purposes, we will not be using the second dataset, although it could yield interesting results as an extensionn of this project.

Prior to analysis, the data will need to be preprocessed by transforming the abundance data into proportional data, as well as removing species/groups with a high proportion of 0. Additionally, the data may require transformation such as a log transformation.

Project Aims

The primary aim of this project is to assess the association between gut microbiome data and ASD patients to see if there are certain bacteria (or bacterial groups) associated more highly with ASD children vs. TD children. After building this model, we would also like to predict whether a child has ASD based on their gut microbiome data.

There exist several obstacles that require attention while analyzing microbiome data. Three of these issues are outlined below:

- Microbiome data exhibits a high-dimensional nature- our dataset comprises of over 1000 species. This high dimensionality necessitates the use of dimensionality reduction techniques to streamline the analysis.
- Microbiome data is over-dispersed, therefore for modeling, the Negative Binomial distribution should be used instead of the Poisson distribution to model the count data
- Microbiome data is zero-inflated. In order to account for this problem, we will use a zero-inflated model for this feature (Xia et al. 2018).

Methods

References (Must be the last section to have bibliography printed after)

Dan, Zhou, Xuhua Mao, Qisha Liu, Mengchen Guo, Yaoyao Zhuang, Zhi Liu, Kun Chen, et al. 2020. "Altered Gut Microbial Profile Is Associated with Abnormal Metabolism Activity of Autism Spectrum Disorder." *Gut Microbes* 11 (5): 1246–67.

Elsabbagh, Mayada, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Carlos Kauchali, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, et al. 2012. "Global Prevalence of Autism and Other Pervasive Developmental Disorders." *Autism Research* 5 (3): 160–79.

Landrigan, Philip J. 2010. "What Causes Autism? Exploring the Environmental Contribution." *Current Opinion in Pediatrics* 22 (2): 219–25.

Lladó Fernández, Salvador, Tomáš Větrovskỳ, and Petr Baldrian. 2019. "The Concept of Operational Taxonomic Units Revisited: Genomes of Bacteria That Are Regarded as Closely Related Are Often Highly Dissimilar." *Folia Microbiologica* 64: 19–23.

Lord, Catherine, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. 2018. "Autism Spectrum Disorder." *The Lancet* 392 (10146): 508–20.

Matson, Johnny L, and Alison M Kozlowski. 2011. "The Increasing Prevalence of Autism Spectrum Disorders." *Research in Autism Spectrum Disorders* 5 (1): 418–25.

Sutcliffe, James S. 2008. "Insights into the Pathogenesis of Autism." *Science* 321 (5886): 208–9.

Tataru, Christine A, and Maude M David. 2020. "Decoding the Language of Microbiomes Using Word-Embedding Techniques, and Applications in Inflammatory Bowel Disease." *PLoS Computational Biology* 16 (5): e1007859.

Vuong, Helen E, and Elaine Y Hsiao. 2017. "Emerging Roles for the Gut Microbiome in Autism Spectrum Disorder." *Biological Psychiatry* 81 (5): 411–23.

Xia, Yinglin, Jun Sun, Ding-Geng Chen, Yinglin Xia, Jun Sun, and Ding-Geng Chen. 2018. "Modeling Zero-Inflated Microbiome Data." *Statistical Analysis of Microbiome Data with R*, 453–96.