The Genetics of ADHD

2023 Supercomputing Challenge

Camila Carreon, Greta Swanson, Luke Rand,

Nandita Ganesan

Final Report

Bioinformatics

Santa Fe Preparatory School

Santa Fe, NM, US

April 5 2023

Contents

1 Executive Summary

Mental health, a widespread issue, warrants more research and regard. Among the vast array of mental disorders, Attention Deficit Hyperactivity Disorder(ADHD), disproportionately affects youth our age, with around 2.8% of adults and 9.4% of youth in the US diagnosed (ADDitude).The symptoms of ADHD include hyperactivity, impulsive behavior, and difficulty paying attention, affecting both learning and daily life. Additionally, many studies demonstrate strong correlations between genetics and ADHD, and likely ADHD is an inherited disorder ("Attention deficit hyperactivity disorder (ADHD) - Causes."). Many issues have arisen within the accuracy of diagnosis of ADHD and the efficiency of analyzing gene sequences without technology that would identify variations in the genomes correlating with ADHD. Several studies have suggested that, "Doctors can misdiagnose ADHD in children due to their age"(ADHD Misdiagnosis). Others have highlighted the similarities of symptoms between bipolar disorder and ADHD, revealing yet another shortcoming in diagnosis. With numbers and rates of ADHD steadily increasing, accurate diagnosis becomes more pressing. Thus our project, through genetic analysis, has found systematic solutions to these problems. Through gene sequencing technology our program will analyze genetic variations for ADHD and aid in diagnosis.

2 Introduction

The relationship between ADHD and genetics is prevalent and can be utilized to advance diagnosis and understanding of the disorder using bioinformatics. Bioinformatics is a field of the utmost importance because of its role in aiding the management of data in the fields of biology and medicine in order to gain a better understanding of the surrounding world. Within bioinformatics, sci-

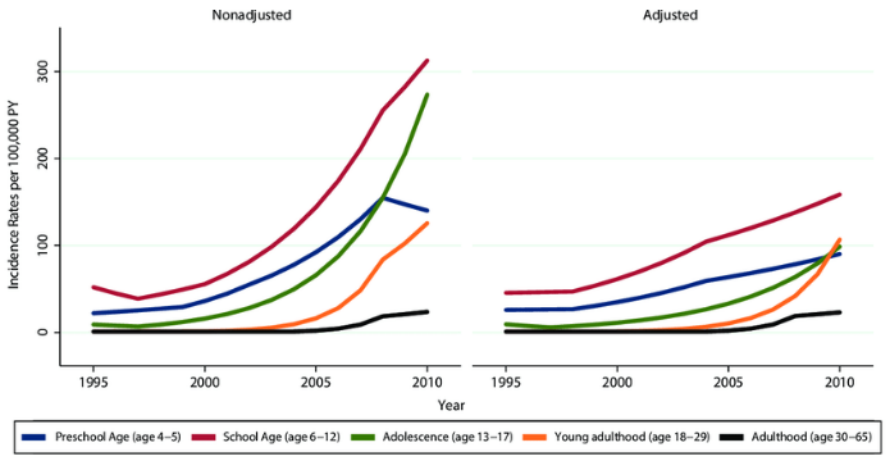

Figure 1: ADHD rates

entists work to analyze gene variation, gene expression, gene structure, protein structure, and gene and protein function (Bayat). Projects such as the Human Genome Project have all been accomplished under the branch of bioinformatics. However, previous work in this field has not been utilized to diagnose mental illnesses correlated to genetics at a large scale. Without an analytical way of diagnosing mental illnesses such as ADHD, doctors' diagnoses remain symptom based. An analytical method to aid the diagnosis of ADHD would allow for a decrease in the number of misdiagnosed patients and an increase in the data surrounding mental illnesses and the specific gene variations they are related to. Our computational model would act as a first step in finding gene variations correlated to ADHD in the diagnosis process, and using this information in conjunction with doctors' assessment of symptoms to provide a strong diagnosis. This information can become additional data to support the advancement of the current research and understanding of ADHD.

3 Background

3.1 DRD4

Although there are many chromosomes linked with ADHD, we have chosen to focus on the variation that occurs on the D4 subtype of the dopamine receptor (DRD4) as the most supported correlation between a chromosome and ADHD is the relationship between DRD4 and ADHD. DRD4 is located on the short arm of chromosome 11 in the position 11p.15.5 and is a protein-coding gene. Typically, within DRD4 variable tandem repeat occurs where there are two or more sequences of 48 sequential base pairs. This variation occurs in exon 3, an exon is part of the gene that forms some of the final RNA, and the variation will have anywhere from 2 to 11 repeats of the base pair. In a study looking at patients with a modification in DRD4-7R (a specific type of modification of the 48 sequential base pairs in 11p.15.5), scientists saw that there was a greater activation of the right temporal lobe in the brain, which deals with processing sensory and emotional stimuli. People with a highly active right temporal lobe are prone to having difficulty paying attention and are often diagnosed with ADHD. The gene variation in DRD4-7R is also a variable tandem repeat modification, with there being 7 repetitions of the 48 sequential base pairs. Additionally, the DRD4-7r variation of the gene is called the "Wanderlust Gene" and is associated with restlessness, and is thus of particular interest when discussing ADHD.

3.2 Privacy in Gathering Data Sets

While, much of our research and programming was smoothly implemented, the issue of privacy in data sharing and access proved a hindrance. The case-control data sets that we pulled from both the GWAS and NCBI catalogues are autho-

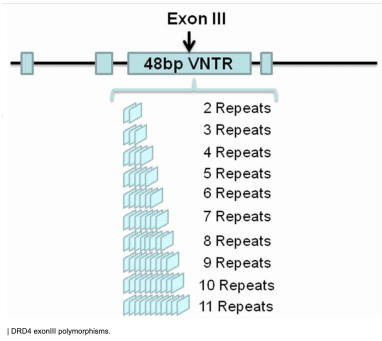

Figure 2: Variations of exon 3 of DRD4

rized for privacy of the patients and volunteers involved in the study. The NIH policy for data management and sharing reads that, "The National Institutes of Health (NIH) is committed to advancing scientific discoveries while safeguarding the interests of study participants and maintaining public trust in biomedical research." Below is a screenshot of the NIH GDS Policy: Thus, we ran into

Part I

Enhancing Consent and Participant Trust through the NIH Genomic Data Sharing Policy

One way to promote trust is to ensure that participants' biospecimens, tissues and cells, and the information derived from them, are used in research only with their permission (consent). Efforts to modernize federal regulations (i.e., the Advanced Notice of Proposed Rule Making to update the Common Rule) are focusing on the need to enhance research protections and reduce regulatory burden. Some of the proposed reforms apply to research using specimens and data and include important changes to informed consent requirements. NIH supports these reforms because seeking consent is respectful to participants and facilitates sharing of biospecimens and data in order to maximize the public benefits of research.

Figure 3: GDS Policy

several problems on acquiring a suitable data set, and this limited the scope of our algorithm. Thus we have outlined our possible algorithm, and possible ways to combat this lingering issue.

3.3 Markov Chain Monte Carlo Method Background

The origins of Markov Chain Monte Carlo methods, known more commonly by the abbreviated counterpart MCMC methods, come out of the nearby Los

Alamos National Laboratory during the 1940s. Stanislaw Ulam, a scientist at LANL, who immersed in developing the Hydrogen Bomb, also worked extensively in combinatorics. The premise of Ulam's idea was to calculate the probability of winning the classic card game, *Solitaire*. What would stem from a musing, would eventually, with the help of his colleague, John Von Neumann, extend to the broader realm of academia. Today, MCMC methods have a variety of applications, and in lieu with our focus, relates to computational biology. Like so many of the methods and ideas incorporated within our project, MCMC methods are often shrouded in jargon and complexity, and thus it is essential to offer a brief overview of how they work. MCMC methods estimate the parameters of a distribution or sample with several dimensions or parameters, through randomly sampling a *proposal distribution*. A proposal distribution which is described as, "the conditional probability of proposing a state given." The MCMC is different from Monte Carlo methods, since it incorporates the idea of Markov Chains which find the probability of events based on the transition from one event to another. Mathematically, we model these transitions through a transition or stochastic matrix. By intersecting the ideas from Markov Chains and Monte Carlo methods, algorithms can effectively and efficiently approximate and estimate parameters from very high dimensional distributions. Genetic distributions pose such problems of high dimensionality, and can thus be addressed through MCMC methods. Fifty years after Stanislaw Ulam and John Von Neumann's work, burgeoning computational and statistical research ushered a revival of MCMC methods and the birth of two major algorithms, Gibbs sampling and the Metropolis Hastings algorithms. In a computational biological context, our group deemed Gibbs Sampling the more appropriate solution. In our project, Gibbs sampling was used in an algorithm to analyze and sample from a data set containing cases and controls and their respective

genetic observations, marking the link between phenotype(the expressed trait) and genotype(the genetic makeup of the individual patients).

3.4 Gibbs Sampling Algorithm Background

Gibbs Sampling, which stemmed from the computer revolution in 1984 was developed and outlined by mathematicians Stuart and Donald Geman. Named in honor of the late J.W. Gibbs, a scientist who worked extensively in thermodynamics, the Gibbs Sampler is an astounding MCMC method. In bioinformatics, the applications of Gibbs Sampling trace back to 1993, when a Gibbs Sampling Algorithm was employed in multiple sequence alignment, which helped identify similarities and differences between multiple(as generally most alignment software only compares two sequences) sequences, but since, there have been many more applications. One is motif finding, where the Gibbs Sampler looks for patterns between sequences of base pairs(A which corresponds to Adenine, C or Cytosine, G which maps to Guanine and T which maps to Thymine) Transcription Factor Binding Sites (TBFS), or the places in the DNA where the process of transcription, where the DNA becomes copied to DNA and ultimately expressed as a protein. In our algorithm, we look at a different application, in which Gibbs Sampling in tandem with an Association Rule Mining algorithm look for patterns in DNA sequences from a ADHD case-control data set.

3.4.1 Gibbs Sampling Overview for the Algorithm

To understand the Gibbs Sampler and it's purpose for our algorithm, educational doctrine suggests using examples. Thus before outlining the Gibbs Sampling application in our algorithm, we will offer a brief and analytical discussion on the Gibbs Sampler: Given a probability distribution that is intractable or extremely hard to sample from $P(x, y)$ And given that the probability distribu-

tions below are very easy to sample from $P(x|y)$ and $P(y|x)$ Then the Gibbs Sampler can be used, where x and y are set to random initial values. A simple Gibbs Sampling Algorithm can be applied where (x_0, y_0) are set to a random starting value. Then to generate then next pair (x_1, y_1) we must sample

$$x_1 \sim p(x|y_0) \tag{1}$$

from the conditional distribution $X|Y = y_0$ to get (x_1, y_0) and subsequently we must sample

$$y_1 \sim p(y|x_1) \tag{2}$$

from the conditional distribution $Y|X = x_1$ to get the final coordinates (x_1, y_1) we can repeat this process some N times where the distribution (x_i, y_i) is dependent upon (x_{i-1}, y_{i-1}). To quantify how well the Gibbs Sampling Algorithm along with all MCMC methods, went, and which will be important for the validation of our algorithm can be modeled by the following function:

$$\frac{1}{N}\sum_{i=1}^{N} h(X_i, Y_i) \tag{3}$$

The Gibbs Sampling Algorithm is a special case of the Metropolis-Hastings Algorithm, another MCMC method. In the Metropolis Hastings Algorithm, new states which are sampled are accepted with a certain probability that can be modeled by the following function:

$$\alpha(\theta^*|\theta) = min\{1, \frac{P(\theta^*|X)Q(\theta)}{P(\theta|X)Q(\theta)}\} \tag{4}$$

In Gibbs Sampling, all states are accepted with probability 1 instead of the above equation. In other words, all proposed states are accepted in Gibbs Sampling. We regard the Gibbs Sampling Algorithm of highest importance

since the data sets we are using (pulled from the GWAS and NCBI Databases) are quite intractable or very hard to sample from, since they contain many parameters. Thus a Gibbs Sampling Algorithm, which dissolves this issue, by sampling from conditional distributions to then estimate a very complex joint probability distribution like the one we are using. The application of Gibbs Sampling in our algorithm is to use Gibbs Sampling in a simulated annealing. In other words, we will be solving an optimization problem in which we are looking for the strongest association rules between SNPs or Single Nucleotide Polymorphisms in an ADHD case control data set[1]. Specifically, the Gibbs Sampling algorithm simulated annealing approach will be used, since the data set we are sampling from is discrete, and generally simulated annealing is used for discrete data sets.

3.5 Association Rule Mining

Association Rule Mining comes from a very different branch of academia, as it was first used by computer scientists Rakesh Agrawal, Arun Swami and Tomasz Imieliński for supermarket transaction data sets. Fundamentally, however, Association Rule Mining looks for relationships between two item sets. Specifically we classify these relationships as "if-then" which are called the antecedents and consequents respectively. In the algorithm we are using, the Association Rule Mining algorithm would look for if then relationships between SNPs and correlation to ADHD. Association algorithms look for the frequency and patterns between these relationships, and use two factors called the support and confidence to determine how important or how prevalent these rules are. In a bioinformatics context, we would be finding the most important SNPs related to ADHD. Association Rule Mining, like MCMC methods has several algorithms

[1] This will be explained in much greater detail throughout the paper, specifically see section 3.4

available to implement it. These include, the Apriori Algorithm, the Eclat algorithm, and the FP-Growth algorithm. Our final project will use the Apriori Algorithm, simply because there is the most research with it in a bioinformatics context. However, to understand the final algorithm in earnest requires an analytic discussion of the mathematical models behind the Apriori Algorithm. First we define

$$Support = \frac{freq(A, B)}{N} \tag{5}$$

Where A and B are items in the item sets. Additionally, we define

$$Confidence = \frac{freq(A, B)}{freq(A)} \tag{6}$$

Finally we define

$$Lift = \frac{Support}{Supp(A) \cdot Supp(B)} \tag{7}$$

By calculating the support, confidence, and lift, the Apriori Algorithm looks for frequent item sets and then develops association rules between item sets in a data set, and follows a rather simple process. First, it calculates the frequency of an item set with 1 item (join step) , then the candidates that meet a minimum support value are chosen, and move forward in the algorithm, while the others are pruned (this is called the prune step). Next, item sets with 2 items are chosen, so for instance if there were 5 items in the 1-item set step, now, there would be $\binom{5}{2} = 10$ items, and the item sets that don't meet the minimum support are pruned. This process is run for several iterations, until at a certain point, the anti monotone rule of the Apriori Algorithm is employed. This rule allows for the pruned subsets to determine what supersets will be pruned. For example, if the item set $\{I1, I2\}$ is not frequent, since it is a subset of $\{I1, I2, I3\}$, the latter item set will be pruned. Then the confidence is found for the remaining item sets, wherein association rules are generated, for instance in the item set

$\{I1, I2, I4\}$, the rule $\{I1, I2\} \geq \{I4\}$. The confidence threshold is then used to choose and prune for the strongest association rules. Additionally, the strength of a rule can be calculated by the lift as shown in formula 7.

4 A General Overview of the Final Algorithm

The Final Algorithm borrowed ideas from both association rule mining and Gibbs sampling to accurately and efficiently choose the SNPs(on DRD4) with the most correlation to ADHD, and thus become a predictor of the disorder in future patients. The first step in employing the algorithm was changing the data set into a "transaction data set" or such that the SNP variables correspond to three different genotypes (0,1, and 2 as outlined in the previous section) and act as predictor variables, while the response variables are modeled by ADHD and NO ADHD (A and NA). The next step in employing the algorithm was to convert this transaction data set into a binary data set, which was done through the arules package in R. Subsequently, the predictor items, which were the SNPs, were renamed as $A_1, ..., A_{750}$. Then an Apriori Algorithm was used on the data set to look for association rules. A random sample of 250 association rules was taken(where the Gibbs Sampling applies), and the frequency in increments of $N = 50$ was calculated. The Apriori Algorithm was then used again to mine for more important association rules. To make use of the Apriori Algorithm for sufficient data mining, we had to choose the minimum support and confidence threshold for the data set. Then by applying a Gibbs Sampling technique to sample the data and then run the Apriori Algorithm multiple times, we were able to select the 15 most prominent association rules. The rules that showed most correlation to ADHD, or the item sets that showed the strongest correlation to ADHD (SNPs on DRD4) would then be chosen as genetic markers for ADHD, and used to assist in ADHD diagnosis. In essence, our algorithm, using an

Apriori algorithm that is improved by Gibbs sampling, would function faster and more accurately on larger data sets (like the one we had in mind), and hence be an accurate tool for doctors and physicians in diagnosing the disorder.

5 An Issue

The Apriori Algorithm, which is generally used in market basket analysis, works best on certain forms of data sets or "transaction data sets". Which look like the following figure 4. Thus in the context of bioinformatics, to apply such an

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Figure 4: Transaction Dataset

algorithm would rely mainly on finding the correct kind of data set to run our algorithm on. The data set we would want looks something like the following figure 5. However, since the molecular data for each volunteer in a case control study was unavailable, because of the privacy concerns outlined in section 3.2,

SNP	Variable	ADHD
rs10499040	1	A
rs2575040	1	NA
rs1804020	1	A
rs7030405	2	A
rs2235040	2	A
rs2280408	2	NA
rs13019040	2	A
rs210405	0	A
rs10040790	1	A
rs7040038	0	NA
rs17040958	0	A

Figure 5: Data

we were unable to acquire such data.

6 Ways to Move Forward

Since there are several privacy issues around acquiring the correct data set, we have highlighted some other approaches that might work instead, but given the limiting time frame, could not be executed. One possible approach, would be to download the nucleotide sequence for SNPs on the DRD4 gene, or SNPs on several sets of genes related to ADHD, and look for patterns between motifs (strings of nucleotides), using the Apriori Algorithm. Additionally, if time permits, another solution would be to employ a random forest algorithm and hamming edit distance techniques to then choose the most prominent SNP biomarkers for ADHD.

7 A note

Since, we weren't able to get the correct data set that would work for the proposed algorithm, we do not have any results or working code for this project.

8 Achievements

While we weren't able to get a working algorithm, one of our greatest achievements was learning, and exploring the intersection of programming and bioinformatics in earnest. Additionally, we learned the importance of teamwork and effective communication. Thus, while never executing our algorithm, the supercomputing challenge proved a medium for growth and learning.

9 Acknowledgments

We would like to thank our mentors Juergen Eckhert and Quinton Flores for their help with our research and feedback throughout our project. We are also very grateful to our sponsor Ms. Comstock for her support and advice

10 Works Cited

References

[1] '1815−Gene Result DRD4 dopamine receptor D4 [(human)]." NCBI, 29 March 2023, https://www.ncbi.nlm.nih.gov/gene/1815. Accessed 6 April 2023.

[2] "ADHD Misdiagnosis: Why Might It Happen?" Medical News Today, MediLexicon International, https://www.medicalnewstoday.com/articles/325595: :text=Doctors%20can%20mis diagnose%2n.d.HD%20in,approximate%2020%25%20difference%20in%20age.

[3] "Attention deficit hyperactivity disorder (ADHD) - Causes." NHS, https://www.nhs.uk/conditions/attention-deficit-hyperactivity-disorder-adhd/causes/. Accessed 9 January 2023.

[4] Bayat, Ardeshir. "Science, medicine, and the future: Bioinformatics - PMC." NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1122955/. Accessed 19 March 2023.

[5] "Blast: Basic Local Alignment Search Tool." National Center for Biotechnology Information, U.S. National Library of Medicine, https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[6] Demontis, Ditte et al. "Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder." Nature genetics vol. 51,1 (2019): 63-75. doi:10.1038/s41588-018-0269-7

[7] "DRD4 Gene." GeneCards, 10 January 2023,https://www.genecards.org/cgi-bin/carddisp.pl?gene=DRD4. Accessed 3 March 2023.

[8] "DRD4 and DAT1 in ADHD: Functional neurobiology to pharmacogenetics."NCBI,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3513209/.Accessed 6 April 2023.

[9] Editors, ADDitude. "ADHD Statistics: New Add Facts and Research." ADDitude, ADDitude, 13 July 2022, https://www.additudemag.com/statistics-of-adhd/.

[10] Faraone, Stephen, et al. "The First Robust Genetic Markers for ADHD Are Reported." Brain and Behavior Research Foundation —, 11 July 2019, https://www.bbrfoundation.org/content/first-robust-genetic-markers- adhd-are-reported. Accessed 9 January 2023.

[11] "Increased brain activity to unpleasant stimuli in individuals with the 7R allele of the DRD4 gene." NCBI, 4 November 2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4272659/. Accessed 6 April 2023.

[12] Kniffin, Cassandra, and Victor McKusick. "Entry - 143465 - ATTENTION DEFICIT-HYPERACTIVITY DISORDER; ADHD." OMIM, 26 February 2013, https://www.omim.org/entry/143465phenotypeMap. Accessed 9 January 2023.

[13] Boosting Association Rule Mining in Large Datasets via Gibbs ... - PNAS. https://www.pnas.org/doi/pdf/10.1073/pnas.1604553113.

[14] "Apriori Algorithm in Data Mining: Implementation with Examples." Software Testing Help, 25 Mar. 2023, https://www.softwaretestinghelp.com/apriori-algorithm/.

[15] "Gibbs Sampling." Wikipedia, Wikimedia Foundation, 26 Mar. 2023, https://en.wikipedia.org/wiki/Gibbssampling.