Contents

Abstract 5

Introduction 6

Research Question 7

Hypothesis 7

Literature Review 8

- Prominence and Importance of Microbiome Research 8
- Gut Microbiota, Microbes, and Metabolic Pathways 8
- Impact of Microbial Metabolism on Host Health 11
- Positive Effects of the Gut Microbiota on Host Health 11
- Negative Effects of Microbial Metabolism on Host Health 11
- Bioinformatics and the Gut Microbiota 12
  - Bioinformatics in Clinical Research 12
  - Genome Sequencing 12
- Association Rules and Association Algorithms 13
  - Association Rules 13
  - Association algorithms - apriori 13
- Examples of association algorithms used in bioinformatical research 14
- Motivation 14

Methods and Materials 15

- Data used in the research 16
- Pre-proccessing 16
- Apriori algorithm 17
- Association rules 17
- Timeline of our work 17

Results 20

- Data description 20
- Pipeline for association rule mining 22
- Association rule parameters dramatically affect the number of generated rules and computational runtime 24
- Identifying non-trivial sets of bacterial metabolic activities 24
- Examples of non-trivial sets 25

Discussion 26

Bibliography 28

Appendices 31

Abstract

The human gut microbiome is of high importance in host's health. Disturbed microbiomes have shown to be significantly associated with proliferation of cancer, increased chances of contracting obesity and type-2 diabetes, weakened immune systems and other diverse medical conditions. In this study, we apply association rule mining to a database of gut bacteria genes, to identify sets of genes that are often jointly present in bacterial genomes. To this end we use data from KEGG database and a recently published collection of microbial genomes. We provide a proof of concept that association rule mining can be effectively utilized to study the human microbiome, and show that some of the identified rules are potentially of biological significance. Specifically, several association rules that we identified may help characterize previously uncharacterized gene families highly common in gut bacteria, and possibly also suggest a role for them in the development of colorectal cancer.

Introduction

The gut microbiota is an enormous collection of microorganisms (also called microbes), which consist mainly of bacteria, archaea, bacteriophages, eukaryotic viruses and fungi existing communally in the gut area. The collective genomes of all these microbes, also known as the gut microbiome, is considerably larger than the human genome – and as such is significantly more complex and difficult to decipher (Lynch & Pedersen, 2016).

Many recent studies have shown strong correlation between human health and the gut microbiota. Moreover, it's been shown that an "unhealthy" gut microbiota – a gut microbiota with low diversity – can cause pathogenesis of various metabolic illnesses, such as obesity and type-2 diabetes (Fan & Pedersen, 2021). The microbiota serves many additional roles in the human body – it's a key factor in development of host immune system during infancy (Fulde & Hornef, 2014); Is key in prevention of pathogen colonization (Kamada et al., 2013); Correlates to pathogenesis of colorectal cancer (CRC) (Ijssennagger et al., 2015); Affects the endocrine system, which is important in both hormonal and behavioral regulation (Neuman et al., 2015). It also helps biosynthesize serotonin (Yano et al., 2015), and affects the host's body adiposity and increases bone density and development (Cho et al., 2012).

The challenging part is using this knowledge to suggest new therapeutic approaches – there is only limited mechanistic understanding, which is an important prerequisite for treatment development. To try and improve said understanding, we explore sets of KEGG orthologs (gene families as coded in the KEGG database) and their relevant metabolic pathways using association rule mining. An association rule mining algorithm can analyze large amounts of data, and derives statistical anecdotes from within it of the form LHS (left hand side of the equation) → RHS (right hand side of the equation) are a disjoint set of items (in our case, KEGG orthologs(KOs)) and the arrow means that the existence of a certain set of KOs on the left hand side implies the existence of another set of KOs on the right hand side.

In our case, we utilized hyperedge sets – a certain type of association rule that instead of the normal LHS → RHS, we have an itemset with a certain amount of items A and any partition of this itemset into LHS → RHS rules passes our constraints for the association rule parameters. Basically, if we split A into any 2 disjoint groups of items B, C then we'll find that B ⇔ C. These association rules can find interesting phenomenon in the microbiome without prior biological knowledge – only based on a statistical analysis of the data. Thus, we can advance our knowledge in a vast range and not focus on only one small part of the microbiome.

Research Question

Are there non-related groups of gene families (KEGG orthologs) that consistently appear in tandem in multiple different gut bacteria? Can we find such gene sets that are currently unknown?

Hypothesis

We hypothesize that there are KEGG orthologs (KOs) that often appear in tandem and yet that these links are currently unknown. There are many known sets of co-occuring genes, such as those taking part in known metabolic pathways, but we focus on non-trivial sets and explore their potential biological importance. Additionally, previous research has shown that mining association rules in general is an effective method to finding groups of itemsets that appear in tandem for biological purposes (Becquet et al., 2002), and we believe that the same goes for KOs.

Literature Review

Prominence and Importance of Microbiome Research

Research of the microbiome has become much more prominent over the past decade, as consequence of technological advances as well as multiple studies showing a strong causal relationship between the health of the human host and their gut microbiota. The gut microbiota serves as a key factor in growth of the host immune system during post-natal and early infant periods, which determines immune system effectiveness in later life (Fulde & Hornef, 2014); Is key in prevention of pathogen colonization (Kamada et al., 2013); Shows high level of importance in pathogenesis of colorectal cancer, as it affects cell proliferation (Ijssennagger et al., 2015); Influences vascular remodeling in the small Intestine (Reinhardt et al., 2012); Regulates functions of the endocrine system, which could bring about new treatments for diseases and disorders affected by hormones, as well as behavioral regulation (Neuman et al., 2015). The gut microbiota also helps regulate and biosynthesize 5-HT (serotonin), which is an important brain neurotransmitter (Yano et al., 2015). It also affects the host's body adiposity and increases bone density and development (Cho et al., 2012). Furthermore, additional studies have shown that an aberrant gut microbiota can cause pathogenesis of metabolic disorders such as obesity and type-2 diabetes (Fan & Pedersen, 2021).

Gut Microbiota, Microbes, and Metabolic Pathways

The gut microbiota is a huge cluster of trillions of microorganisms, also known as microbes, consisting of mainly bacteria, but also including archaea, bacteriophages, fungi and more. These microbes' collective genome is many times larger and more diverse than the human genome and as such is much more complex and difficult to comprehend. Many of these microbes rarely exist isolated. Usually, they are parts of complex, interlinked microbial communities, and are commensal – meaning that they benefit from each other and from the host, without hurting it. The more we progress in research of microbes, the more we can see that almost every single creature on earth has a unique collection of microbes that serve as its microbiota (Lynch & Pedersen, 2016). These microbes are metabolically active, meaning that they constantly degrade, consume, and produce metabolites (small molecules) in the gut environment (Downs, 2006). See figure 1 for an example of some important metabolites - Short-chain fatty acids (SCFA). This metabolic activity can often be sorted and organized into "metabolic pathways" (see figure 2 for an example) – series of chemical reactions occurring in a certain order and serving a specific function (Nelson, 2008). Each metabolic pathway consists of dozens, if not hundreds, of metabolites – and each metabolite takes part in many metabolic pathways. As consequence of this, the metabolic pathways are intertwined and create an intricate and detailed web of metabolites and chemical processes. There are several online databases that include lists of metabolic pathways and their associated metabolites and chemical reactions, such as the KEGG database (Kanehisa & Goto, 2000) or MetaCyc (Caspi et al., 2020).

Figure 1: SCFA's and their importance in host health (Śliżewska et al., 2021)

Short-chain fatty acids are small fatty acids that have fewer than 6 carbon atoms. They are produced by microbes in the host's intestines when nondigestible carbohydrates, undigested food fibers or resistant starch get fermented, and have important roles in host health (Śliżewska et al., 2021). They show importance in gastrointestinal physiological health, the host immune system, and host metabolism. They also seem to be important in various cognitive functions, including Alzheimer's disease and Parkinson's disease (Silva et al., 2020).

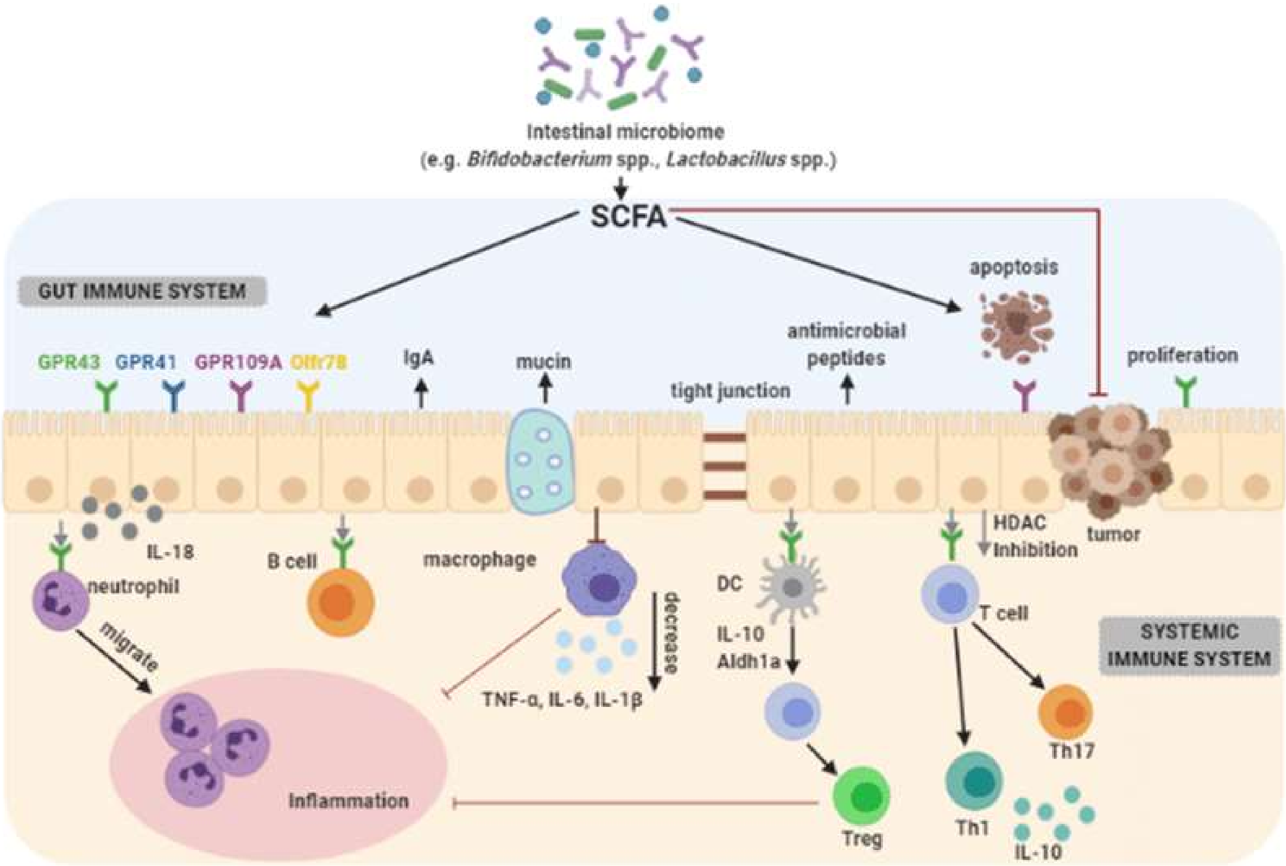
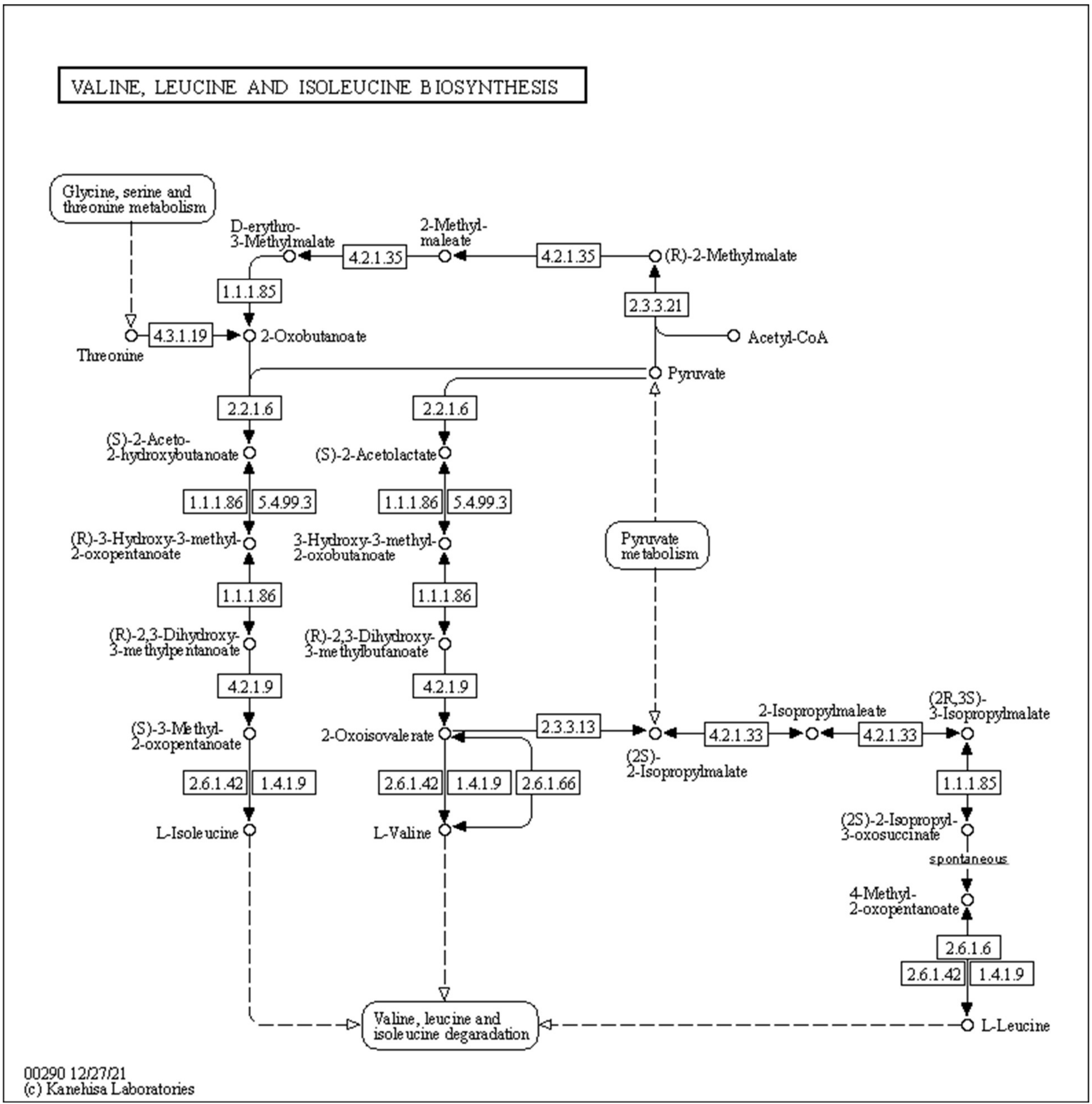


Figure 2: Valine, Leucine and Isoleucine biosynthesis metabolic pathway



00290 12/27/21
(c) Kanehisa Laboratories

This figure is an example of a metabolic pathway, taken from the online KEGG database. The metabolic pathway shown in this image is a pathway that biosynthesizes the branched-chain amino acids valine, leucine and isoleucine. Each node on the map represents a metabolite, and each directed edge represents a chemical reaction, with the rectangles showing codes that represent the enzymes causing said reaction. The meaning behind the edges being directed is that a chemical reaction can only occur in the "direction" the edge goes and cannot necessarily be reversed in the same fashion (Kanehisa et al., 2017).

Impact of Microbial Metabolism on Host Health

The metabolic activity of microbes in the gut has both negative and positive impact on the host's health, and we must research these connections between microbes and metabolites to truly understand the mechanics behind the microbiota's metabolic activity (Fan & Pedersen, 2021).

The microbes residing in the human gut are known to contribute to host metabolism. Yet, over millions of years, the microbial communities hosted in our human body have evolved and adapted to the environment, just as we have. This has led to interdependence – meaning, the microbes rely on us to survive, and we in turn depend on them for many more health functions, beyond metabolism (Lynch & Pedersen, 2016). The importance of each of these roles differs from person to person, but it's safe to say that the microbiome does indeed hold a large stake in host health (Lynch & Pedersen, 2016).

Positive Effects of the Gut Microbiota on Host Health

One example of an important health benefit of the human gut microbiome is related to development of the host immune system during postnatally and early infant periods, which is highly important when determining immune system effectiveness in later life. Before birth, the host's intestinal tract houses very little, if any, microbial organisms. Upon birth, and in the very early stages of infancy, the host is exposed to many microbial organisms at once from their environment. Method of delivery, hygiene conditions at birth, as well as microbial and environmental exposure and other factors all seem to affect early development, and effectiveness, of the host immune system (Fulde & Hornef, 2014).

Negative Effects of Microbial Metabolism on Host Health

There are also examples of microbial metabolism negatively affecting host health. One well-documented example is the case of how consumption of red and processed meat affects cardiovascular risk in the host. The consensus used to be that the fats and cholesterol levels (Bernstein et al., 2010) or high sodium levels (Micha et al., 2010) in meat were the main risk factor of cardiovascular disease caused by red meat consumption, but recent findings seem to point in a different direction – namely, microbial metabolism. Intestinal microbes metabolize choline and phosphatidylcholine, which produces trimethylamine (TMA) which is then oxidized, creating trimethylamine-N-oxide (TMAO). TMAO production has been shown to have strong correlation with higher risk of cardiovascular diseases. TMAO is also produced by metabolizing L-carnitine, a common nutrient in meat. Thus, consuming meat seems to be directly tied towards production of TMAO and inadvertently, increases risk of cardiovascular disease (Koeth et al., 2013).

Clinical Interventions?

While clinical interventions are a long-term goal, manipulating the gut microbiota to improve health outcomes has only been successful in very few use-cases. In order to advance microbiome-based therapeutics, a mechanistic understanding of how the microbiota interacts with its environment and the host must be achieved. One invaluable approach for studying the gut microbiota is through bioinformatic (computational) analysis.

Bioinformatics and the Gut Microbiota

Bioinformatics in Clinical Research

Bioinformatics is a scientific field that utilizes advanced software tools to analyze and better understand biological data on a large scale. It combines many different fields of science, including computer science, biology, chemistry, and statistics. One of the main focuses of bioinformatic research is the analysis of whole genome sequencing (WGS) data (Lesk, 2002). WGS is the act of sequencing the entire genome of a chosen organism. The most famous example of this is The Human Genome Project, a research project of massive proportions that ultimately led to an infinitely deeper understanding of our bodies (Collins et al., 2003). Genome sequencing is a main factor in our research as well – by studying the sequenced genomes of gut bacteria, with bioinformatic methods, we hope to achieve a deeper understanding of these bacteria's metabolic functions.

Genome Sequencing

In recent years, there have been massive developments in sequencing technologies. These technologies have brought about a rise in bioinformatic-based research, as there is a lot more data to work with.

The first method for genetic sequencing was invented in 1975 by Frederick Sanger. In 1977, two more important papers were published – one by Sanger and his colleagues, which described a new chain-termination approach to DNA sequencing (Sanger et al., 1977), and one by Allan Maxam and Walter Gilbert that breaks apart the DNA molecule to sequence it (Maxam & Gilbert, 1977). These two approaches were the baseline for the development of significantly more advanced technologies. The current method for genome sequencing is called "next generation sequencing" (NGS). NGS allows for extremely high throughput by sequencing millions of fragments of DNA in parallel (Behjati & Tarpey, 2013). In microbiome research, NGS is combined with shotgun sequencing to produce "metagenomic" data. This data consists of short DNA sequences originating from all genomes present in a sample, which can consist of many millions of different organisms.

Association Rules and Association Algorithms

Association Rules

Association rules are the basis of the association algorithm, which we will be using to analyze the genetic data from the microbiome. Association rules in a practical sense allow you to "connect" between items from a set, and say that the existence of one item (or a group of items) would imply the existence of another – for example, if the set is a possible shopping list, an association rule could be Milk → Eggs, meaning that if someone has milk on their shopping list, there's a high chance they'll have eggs as well.

Association rules are strictly defined as such: "LHS → RHS", where LHS (*left hand side* of the equation) and RHS (*right hand side* of the equation) are disjoint sets of items (e.g genes or metabolites), and the arrow means that the existence of LHS implies the existence of RHS. In other words, if the LHS set exists in a sample, then the RHS set has a high change of existing as well (Anandhavalli et al., 2010). Association rule algorithms can find connections between metabolites, without need of prior biological knowledge – just by analyzing existing data.

Association algorithms - apriori

"Apriori" is a relatively basic algorithm meant to identify "frequent itemsets" from a list of transactions. A "frequent itemset" is an itemset that is included in more than a certain percent of transactions (the exact threshold is configurable). Apriori uses the fact that if some small itemset is not frequent, then no larger set containing this itemset will be frequent (hence – no need to explore it). Based on these frequent itemsets, association rules can be extracted.

Instead of trying to find association rules in an entire set of data, the "apriori" algorithm divides the data into subsets – small groups of the data. If a small subset is common enough, Apriori then tests it for association rules. Let's take a shopping list as an example: bread and milk are very common items, so apriori will look for association rules such as Milk → Bread and Bread → Milk. If a certain subset of data isn't common enough on its own, there's no reason to test for association rules in it – and that's what apriori utilizes to reduce runtime complexity.

Another example can be with metabolites. Let's say we have a common subset {A,B,C,D} where each element in the subset is a metabolite. Apriori will check for association rules within said subset, for example: {A,B,C} → {D}. If, for example, we have another subset: {A,B,C,D,E} that is no longer common enough in our data, apriori won't bother testing it – so the association rule for the equation {A,B,C,D} → {E} won't ever be calculated, thus saving time.

Examples of association algorithms used in bioinformatical research

One example of mining association rules to provide biological insight is mining association rules in gene expression data. Gene expression data can easily be portrayed in a matrix, and analyzing a complex matrix is precisely the task association algorithms are suited for. The goal of the research was to test whether generated association rules could point towards biologically interesting phenomena, and the results showed that indeed, even though the algorithm had no prior biological knowledge, biologically relevant patterns were picked up. An example of this is that there are some ribosomal mRNAs that are strongly co-expressed (Becquet et al., 2002).

Another example is very similar to the work done in this research. The researchers used an upgraded and more advanced version of the association rule algorithm to mine association rules regarding co-occurrence of human-associated microbial species (Liu et al., 2021).

Motivation

Our motivation for this research is to generally increase our understanding of the human gut microbiome using advanced data analysis methods. We hope to find non-trivial groups of genes with an association rule mining approach, that could potentially lead to the definition of new metabolic modules and elucidate the complex functions of bacteria in the human gut.

Methods and Materials

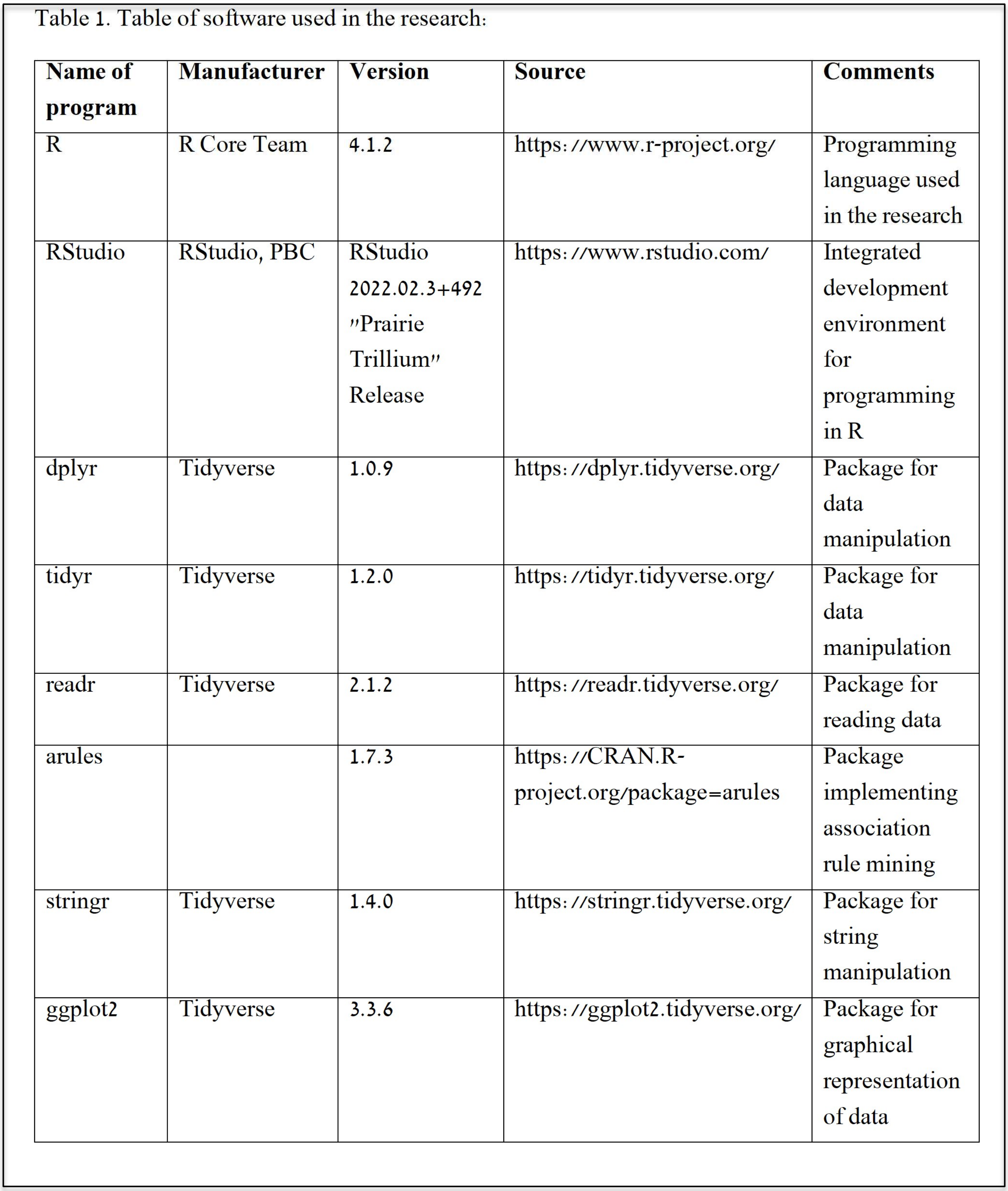


Table 1. Table of software used in the research:

Name of program	Manufacturer	Version	Source	Comments
R	R Core Team	4.1.2	https://www.r-project.org/	Programming language used in the research
RStudio	RStudio, PBC	RStudio 2022.02.3+492 "Prairie Trillium" Release	https://www.rstudio.com/	Integrated development environment for programming in R
dplyr	Tidyverse	1.0.9	https://dplyr.tidyverse.org/	Package for data manipulation
tidyr	Tidyverse	1.2.0	https://tidyr.tidyverse.org/	Package for data manipulation
readr	Tidyverse	2.1.2	https://readr.tidyverse.org/	Package for reading data
arules		1.7.3	https://CRAN.R-project.org/package=arules	Package implementing association rule mining
stringr	Tidyverse	1.4.0	https://stringr.tidyverse.org/	Package for string manipulation
ggplot2	Tidyverse	3.3.6	https://ggplot2.tidyverse.org/	Package for graphical representation of data

Data used in the research

gene_mappings.csv: Table consisting of 6 columns – gene, genome, bac_name, KEGG_gene, ko, gene_weight. Created by scraping and organizing data from a paper which catalogued over 200,000 gut microbiome genomes (Almeida et al., 2021b).The table was prepared in advance by the Borenstein lab.

This dataset consists of 200,000+ different genomes grouped together by similarity. Of each group, one genome was chosen as a representative genome which was further analyzed. Overall, 3,853 representative genomes were analyzed. Next, for each of the 3,853 genomes, the Prodigal program was applied (prodigal is software that can mark genes within a genome sequence, without much prior knowledge about the genes). Each gene marked by Prodigal was then cross referenced with the gene catalog in KEGG database, and the best hit for the gene was chosen for each one. Based on the best hit, each gene was also assigned a KEGG orthology (KO) gene family identifier.

pathway_ko.list: A list of all known metabolic pathways and each of their instances in a different ko in KEGG database. Downloaded from the KEGG website.

The file is a compilation of all documented KEGG orthologies, and all their metabolic pathways. We cross-referenced this data with the data we got from mining association rules, to try and find KEGG orthologies that have correlation – that is, KEGG orthologies that appear together a lot in different genomes, and as such appear together in association rules.

rules_s0.45_c0.8.txt: text file of all the rules we found when running apriori algorithm on the dataset (gene_mappings.csv) when looking for edgesets with the support parameter being 0.45 and the confidence parameter being 0.8.

Pre-proccessing

After obtaining the gene_mappings.csv file, we organized the data into two dataframes - groups by genomes and groups by KOs. The dataframe grouped by genomes now had the genomes as rows and number of KOs per genome in respect to each genome as the data. The dataframe grouped by KOs had KOs as rows and the number of genomes per KO in respect to each KO as the data.

After mapping the KO's to the genomes, we filtered out the rows in our dataframe that weren't distinct or had no KO value, as we are mining association rules between KO's in different genomes and rows without KO values are irrelevant.

Afterwards, we filtered out KOs that occur in above 70% of the genomes as they're most likely ubiquitous functions of bacteria, and won't be relevant to novel results.

On this new dataframe we ran apriori algorithm.

Apriori algorithm

"Apriori" is a relatively basic algorithm meant to identify "frequent itemsets" from a list of transactions. A "frequent itemset" is an itemset that is included in more than a certain percent of transactions (the exact threshold is configurable). Apriori uses the fact that if some small itemset is not frequent, then no larger set containing this itemset will be frequent (hence – no need to explore it). Based on these frequent itemsets, association rules can be extracted.

Association rules

Association rules are strictly defined as such: "LHS → RHS", where LHS (*left hand side* of the equation) and RHS (*right hand side* of the equation) are disjoint sets of items (e.g genes or metabolites), and the arrow means that the existence of LHS implies the existence of RHS. In other words, if the LHS set exists in a sample, then the RHS set has a high change of existing as well. Association rule algorithms can find connections between metabolites, without need of prior biological knowledge – just by analyzing existing data.

Association rules have two parameters – support and confidence. Support is the frequency of an itemset in the dataset. For example, our dataset is genomes and KO's – so an association rule with 0.8 support consists of several KO's such that they appear together in above 80% of the genomes.

The confidence parameter refers to how "confident" the algorithm is that if an item X appears in an itemset, a different item Y will appear as well. For example, say X is one KO and Y is another, if in many genomes both X and Y appear then a rule that satisfies both will have high confidence. This is calculated by dividing the number of genomes who have both X and Y KOs divided by the number of ones that have only X KO.

Timeline of our work

The first thing we did was download all necessary software and data, included in the first table and in the previous page respectively. We are going to analyze large amounts of data using the R language, so we downloaded R and the most widely used integrated development environment (IDE) for R, which is RStudio.

As we had large amounts of data, we downloaded different libraries from the “tidyverse” suite (Wickham et al., 2019) that help with data manipulation and restructuring. Using that, we filtered out irrelevant data from gene_mappings, which is rows with no gene in the KEGG library. These are rows in the table that we cannot analyze using KEGG, so they cannot help us with our research. We then narrowed down the rows to be completely distinct.

Next, we wrote a function that called the apriori function on our data and mined association rules. We saved the appropriate data after before cleaning up irrelevant rules (rules that were subsets of other rules). Table 2 summarizes the results we got from 12 runs of our function, each time with different “Support” and “Confidence” parameters:

Support	Confidence	Runtime	Number of rules	Maximal rules	Largest rule size
0.6	0.8	0.64	38	20	4
0.6	0.85	0.64	38	20	4
0.6	0.9	0.63	32	14	4
0.55	0.8	0.74	370	168	5
0.55	0.85	0.72	315	118	5
0.55	0.9	0.74	226	77	5
0.5	0.8	4.36	2696	848	8
0.5	0.85	2.91	2293	675	8
0.5	0.9	2.56	1985	562	8
0.45	0.8	3.71	28278	7847	9
0.45	0.85	3.60	27412	7653	9
0.45	0.9	2.33	22348	5892	9

Table 3. Table of hyperedge set mining results

There are 6 columns in this table. The first two show the support and confidence parameters for each run of our function. The third column is runtime in seconds. Many things can affect the runtime, such as traffic on the server running the code, but generally we'll see that the more rules mined, the higher the runtime. The fourth column represents the total number of rules mined, including insignificant rules. The fifth column shows the number of maximal rules, which are the largest rule (the rule with most items) X such that there is no rule Y that is larger than X and $X \subset Y$. There are obviously many less maximal rules than the total rules, but there are still thousands in the higher support/confidence ranges. The sixth and final column shows the size of the largest rule. We observe that the lower the support and confidence, the larger the rules are. When we tried to run the algorithm at lower supports, the computation was too heavy for my personal computer and we therefore aborted those runs.

Results

In this research, we mined association rules in gene families associated with the human microbiome. Our research had two goals – the main one was proving that association rules can be used to find novel relations between groups of genes in the microbiome, and in more general terms, that association rules can be used in microbiome research. Our second goal was trying to find a few such novel gene sets and gauge their potential of being biologically meaningful and interesting for future research.

Data description

Our data consists of 204,938 reference genomes grouped together into 3,858 genome clusters by similarity (Almeida et al., 2021a). These could be roughly thought of as different species or sub-species. Each genome was mapped to all genes identified within it, coded using the KEGG Orthology (KO) gene family identifiers.

We then grouped the data in a matrix of genome by KO to see which KO's exist in which genomes and vice versa.

Altogether in the data we had:

- 3,853 genome clusters
- 8,443 distinct KO's
- 3,827,714 total KO's
- An average of 993.44 KO's per genome (see fig. 1)
- An average of 453.36 genomes per KO (see fig.1)

Afterwards, we filtered out the KO's that occurred in over 70% of the genomes, as they most likely represent a group of genes performing a very basic function of the bacteria and as such will overwhelm our association rules with trivial results (see fig. 2).

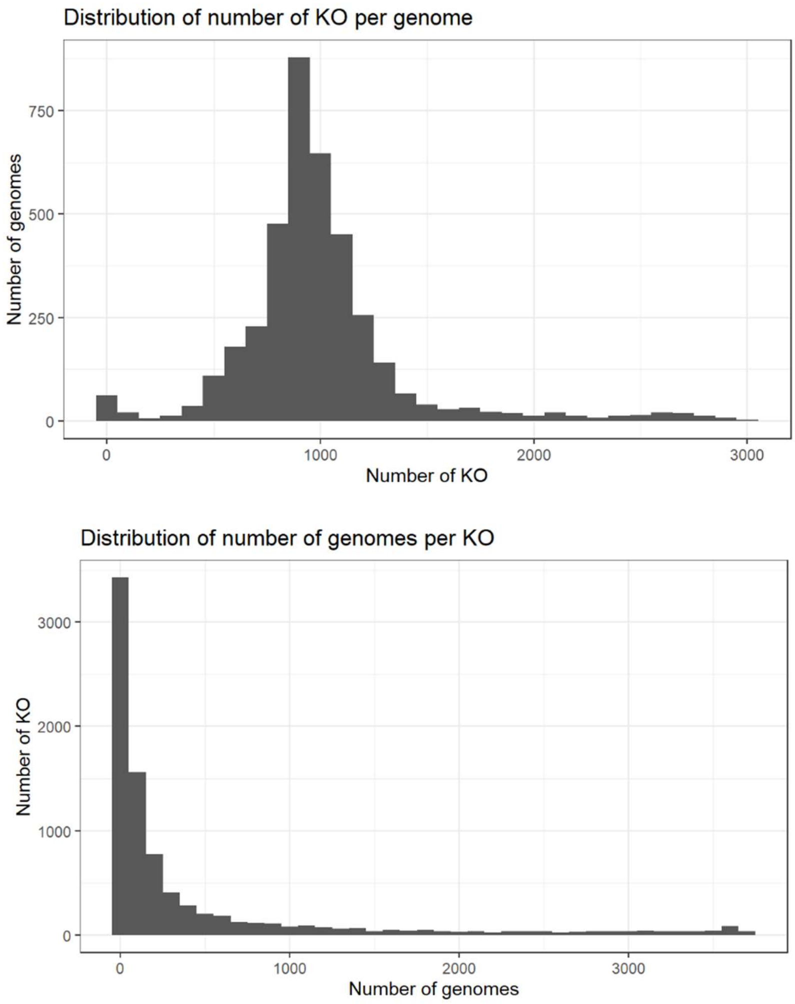


Fig 3. Distributions of KO's and genomes:

These two images portray histograms of distributions of the number of KOs per genome and vice versa. The top histogram shows the amount genomes (y axis) that have a specific amount of Kos (x axis). For example, there are ~1000 KO's that show up in ~600 genomes.

The bottom image portrays the reversed relation – the number of KOs associated with a given number of genomes. For example, there are ~200 genomes in which more than 1,000 of the same KO appear.

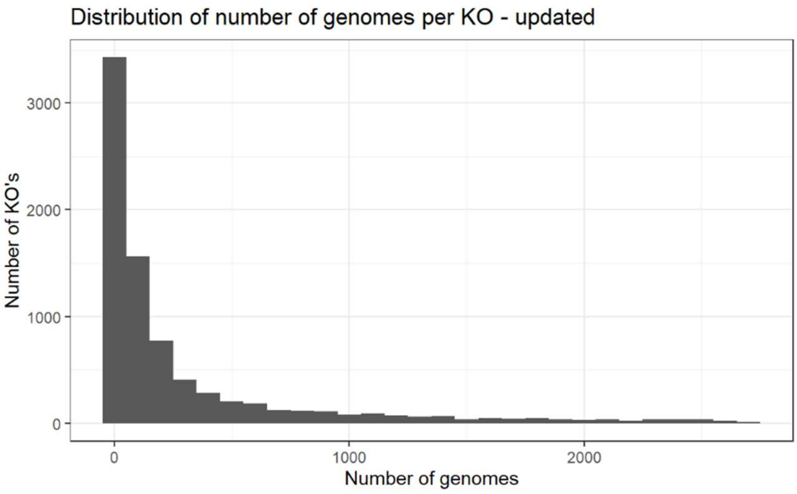


Fig 4. Updated distribution of genomes per KO:

After filtering out the KO's that occurred in over 70% of the data, we have a new histogram, where there are no KO's that appear in over ~2800 genomes.

Pipeline for association rule mining
After having downloaded and organizing the data, we developed an association rule mining pipeline, utilizing the apriori library in the R programming language.

Our first decision was to try and find both hyperedge sets and "normal" association rules. We previously defined what standard association rules are, following the logic: "if X items appear then Y items have a high probability of appearing as well". Hyperedge sets are very similar but require that any partition of the set into 2 subsets of items, namely X and Y, fulfil the following conditions:

- If X items appear then Y items have a high probability of appearing
- If Y items appear then X item have a high probability of appearing

For example: if the hyperedge set is {A,B,C} then {A,B} ⬄ {C}, {A,C} ⬄ {B}, {B,C} ⬄ {A}.

What this means is that we'll find rules in which the association between the items is bidirectional and perhaps stronger, but we'll also find substantially less rules. We eventually decided to mine only hyperedge sets in our research for three reasons:

1. It's more straight forward to filter out trivial rules.
2. Hyperedge sets are more conservative than association rules, acting as an additional filter to identify only the strongest associations.
3. Due to time constraints we decided to focus on hyperedges sets only.

We ran the apriori function with two required parameters, namely "support" and "confidence", and saved results to a data table. We ran our pipeline with multiple support+confidence parameter pairs to explore the landscape of identified association rules before choosing one parameter pair for further analysis of associations. After settling on a single such setting, we next filtered out "irrelevant rules" (see below) and retained only non-trivial rules.

For a comprehensive visual representation of the steps taken in our research see Fig 5.

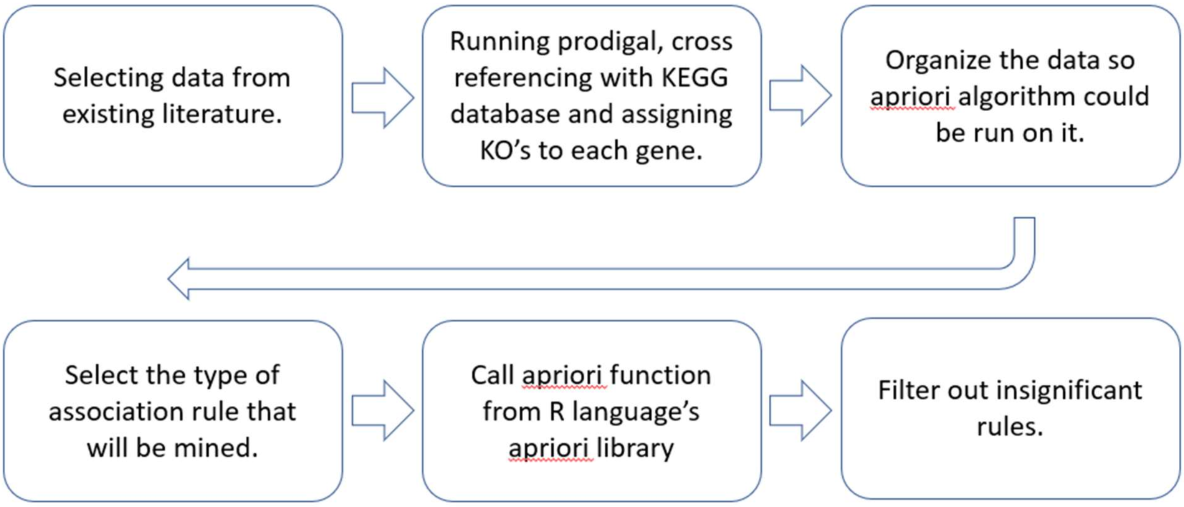

Selecting data from existing literature.
Running prodigal, cross referencing with KEGG database and assigning KO's to each gene.
Organize the data so apriori algorithm could be run on it.
Select the type of association rule that will be mined.
Call apriori function from R language's apriori library
Filter out insignificant rules.

Fig 5. Pipeline for association rule mining:

From left to right, the 6 steps we took in our research to mine association rules.

First, we selected data from existing literature - 204,938 reference genomes grouped together into 3,858 groups by similarity (Almeida et al., 2021a).

Secondly, we ran prodigal software to mark the genes from the data and assign their kegg orthologs.

Thirdly, we organized the data in histograms to better understand it, and then organized it as transactions for the apriori function input.

Fourthly, we selected the type of association rule to be mined. There are multiple factors to this decision, but we selected hyperedge sets.

Next, we called the apriori function from the apriori R library as such:

```
rules <- apriori(gene_mappings_transactions, parameter = list(support = sup, confidence = confd, "target" = "hyperedgesets", maxtime = 0, maxlen = 4000, ext = F))
```

gene_mapping_transactions is the data organized in transactions, which is the input from which the association rules will be mined. "parameter" is a list of the parameters – support and confidence are the selected support and confidence cutoffs, "target" is the type of association rule to be mined – in our case hyperedge sets. "maxtime" is the maximal amount of time the algorithm will run. By setting it to 0, there is no maximal time and the algorithm can run indefinitely. "maxlen" is the longest possible rule, we set it to 4,000 to make sure we won't miss any long rules.

Finally, we filter out trivial rules.

Association rule parameters dramatically affect the number of generated rules and computational runtime

Confidence and support parameters directly affect the number of rules mined. Specifically, the higher you set them, less rules satisfy them (see table 2). If you require a support of 0.6 and a confidence of 0.9, that means that any itemset found by the algorithm must appear in 60% of the dataset and that all items in the rule appear together 90% of the time. These are highly conservative constraints, and as such there will be very few rules that satisfy them – but the runtime will be low, as there will be very little rules to calculate. On the other hand, by setting lower thresholds, such as 0.45 support and 0.8 confidence, an extremely high number of rules is generated, but high runtime as well (see table 2).

Identifying non-trivial sets of bacterial metabolic activities

After mining the association rules using different parameters (see table 2), we decided to further investigate itemsets that were identified using a support level of 0.45 and a confidence level of 0.8. This resulted in thousands of different rules, which we then further narrowed down by applying the following filters:. Our first idea was to calculate p-values for the rules (using a chi-square test) and remove rules with higher p-values, but we quickly realized that due to our conservative parameter choice almost all rules would be statistically significant and using p-values barely narrowed down our results.

We then noticed that there were many rules that were subsets of other rules – and those rules could be removed. For example, if we have a certain rule: $\{A, B, C, D\}$ and another rule $\{A, B, C, D, E\}$, then we can remove the first rule as it is contained in the second, and therefor redundant. With this in mind, we designed an algorithm that removed all rules that were subsets of other rules by sorting the rules from largest to the smallest, and then temporarily saving the largest ones and iterating over the rest of the rules to remove the subsets of said rule. This managed to significantly narrow down our results (see table 2).

After doing that, we started cross-referencing our rules with the known metabolic pathways saved in KEGG database. Each metabolic pathway was viewed as a list of Kos contained within it (as edges in the pathway network). We first removed the 11 global and overview maps in KEGG database, as they're very large pathways that contain over hundreds of chemical reactions not neccasirly reflecting a specific function. We then decided to remove all pathways that had over 50 reactions in them as they are similarly likely to represent multiple functions and not nessarily imply that all KO subsets contained within them are trivial sets. After doing this, we removed all rules that were subsets of the metabolic pathways left, as they're trivial – and focused on the "non-trivial" ones that cannot easily be explained by serving one specific function as those captured by known pathways. By doing this, we narrowed down the number of rules we had to go through by hand to a few hundred, with each one having a relatively high likelihood of being non-trivial in some sense.

Examples of non-trivial sets

We present below two examples of rules we identified and discuss there potential biological significance:

1. {K03091,K06409,K06960}: According to KEGG database, K03091 and K06409 are both genes tied to sporulation in certain bacteria, while K06960 is an undocumented protein. All three of them appear in 500 different bacterial species, but K06960 appears almost exclusively in bacteria in which K03091 and K06409 appear. That means that there's a chance it's tied to sporulation, or another function performed exclusively by bacteria that can perform sporulation. Additionally, K06960 has been found to be predictive of lynch syndrome, which is the most common cause of hereditary colorectal cancer, meaning that in patients in which it appears there is a high likelihood of developing colorectal cancer (Yan et al., 2020). There could also be a connection between that and the other two genes in the set, but more extensive research will need to be done.
2. {K18888,K18887,K16785,K07166,K09157}: According to KEGG database, the first three KO's - {K18888,K18887,K16785} – all appear within a single metabolic pathway, pathway 2010 – which is a pathway of ABC transporters, which are proteins that are responsible for translocating solutions using the energy from ATP hydrolysis (Jones & George, 2004). However, the second two KO's – {K07166,K09157} - aren't uncharacterized in the literature, and we suggest that they might also be related to ABC transporters.

Overall we show that our algorithm can be used to generate hypotheses regarding the function of different uncharacterized genes, and highlight genes that could potentially be related to human diseases.

Discussion

(Almeida et al., 2021b)The human microbiome and specifically the gut microbiome is of high importance to human health, and modern research shows that it is inherently tied to many body functions such as immune system development (Fulde & Hornef, 2014), prevention of pathogen colonization in the gut (Kamada et al., 2013), cell proliferation (Ijssennagger et al., 2015), vascular remodeling (Reinhardt et al., 2012), and functions of the endocrine system and behavioral regulation (Neuman et al., 2015). Gut microbiome composition has also been associated with pathogenesis of different diseases such as obesity and type 2 diabetes (Fan & Pedersen, 2021), as well as colorectal cancer (Ijssennagger et al., 2015).

A main challenge in microbiome research nowadays is the incomprehensible amount of generated data and the complexity of the entire ecosystem, consisting of countless microorganisms interacting with each other, with their host, and with other environmental factors. In this work, we suggested a method to analyze large amounts of microbiome-related data, utilizing association rule mining, and hypothesized that such analysis can potentially lead to novel and non-trivial findings about the function of human gut bacteria.

Our approach was to generate multiple biological hypotheses automatically using association rule mining, narrow down the rules to those that we believe are non-trivial, and then perform further research only on those. One example of what we could find is a group of KEGG orthologies (KO's) that perform some biological function that was not known before – which could help us gain a mechanistic understanding of the microbiome.

Using our algorithm, not only have we narrowed down the immense amount of data, but we also managed to begin working on filtering out non-trivial results. In essence, we managed to narrow down the data even further to raise the probability of our results being both important and novel. One example of such a result is an association rule we found: {K03091,K06409,K06960}. As stated earlier in our results, K03091 and K06409 are both genes related to sporulation (KEGG entry about the genes, insert the research paper here later), and K06960 may be related to lynch syndrome (Yan et al., 2020) – the most common form of hereditary colorectal cancer – meaning researching this rule could even result in important findings about colorectal cancer.

Our research, however, has several limitations. First and foremost was the limited computational power – we never got to run the algorithm with the parameters we had hoped for, due to non-realistic running times. We also faced a tight schedule and left some parts of the algorithm in theirnaïve version, making the computational complexity high. In general, mining association rules using apriori algorithm is a very computational heavy task, so any improvement in the runtime would greatly increase the effectiveness of our research – allowing us to run the algorithm with much lower support values, and potentially uncovering less common but important (and large) association rules.

The strength of our approach to microbiome research is the huge amount of data we can analyze – but reliance on previously uncovered data means that finding truly novel results could be difficult, and all results must be authenticated in a lab in vitro/in vivo. While our research is effective at pointing out interesting phenomenon that could have important biological functions, said phenomenon still need to be studied in a lab to truly gain an understanding of them.

We believe there to be multiple interesting continuations to this research. First of all, narrowing down the KO's on which we run the association rule mining algorithm to the least common 30% would allow us to run on a much lower support parameter, which would inevitably result in larger and potentially more interesting rules. It would also narrow down the association rules to KO's that are less common, which on average means that they're also less well documented – meaning we may find novel results that we otherwise would have missed.

Another possible continuation is doing very similar research on a server with higher computational power, which would allow us to run this exact algorithm but mine association rules with a much lower support parameter, meaning we could find very large rules that could be interesting. In addition, designing the algorithm to also mine normal association rules, and not only hyperedge sets, and see what kind of results we get. We'll obviously get more results, as every hyperedge set will end up being more normal association rules, but we may also find rules that the current algorithm mining hyperedge sets misses, and they might showcase important causal relationships between KO's and chemical reactions.

Alternatively, running the algorithm on different data – instead of mining association rules between KO's based on chemical reactions, we could try and run the algorithm on data about metabolites and bacteria, to try and find which metabolites are synthesized together by the gut microbiota, which could have interesting biological and chemical results.

Bibliography

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021a). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021b). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Anandhavalli, M., Ghose, M. K., & Gauthaman, K. (2010). Association Rule Mining in Genomics. *International Journal of Computer Theory and Engineering*, 269–273. https://doi.org/10.7763/ijcte.2010.v2.151

Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F., & Gandrillon, O. (2002). *Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data.* http://genomebiology.com/2002/3/12/research/0067.1

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, *98*(6), 236–238. https://doi.org/10.1136/archdischild-2013-304340

Bernstein, A. M., Sun, Q., Hu, F. B., Stampfer, M. J., Manson, J. E., & Willett, W. C. (2010). Major dietary protein sources and risk of coronary heart disease in women. *Circulation*, *122*(9), 876–883. https://doi.org/10.1161/CIRCULATIONAHA.109.915165

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., Ong, W. K., Paley, S., Subhraveti, P., & Karp, P. D. (2020). The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic Acids Research*, *48*(D1), D455–D453. https://doi.org/10.1093/nar/gkz862

Cho, I., Yamanishi, S., Cox, L., Methé, B. A., Zavadil, J., Li, K., Gao, Z., Mahana, D., Raju, K., Teitler, I., Li, H., Alekseyenko, A. v., & Blaser, M. J. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature*, *488*(7413), 621–626. https://doi.org/10.1038/nature11400

Collins, F. S., Morgan, M., & Patrinos, A. (2003). *The Human Genome Project: Lessons from Large-Scale Biology*. www.sciencemag.org

Downs, D. M. (2006). Understanding microbial metabolism. In *Annual Review of Microbiology* (Vol. 60, pp. 533–559). https://doi.org/10.1146/annurev.micro.60.080805.142308

Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. In *Nature Reviews Microbiology* (Vol. 19, Issue 1, pp. 55–71). Nature Research. https://doi.org/10.1038/s41579-020-0433-9

Fulde, M., & Hornef, M. W. (2014). *Maturation of the enteric mucosal innate immune system during the postnatal period.*

Ijssennagger, N., Belzer, C., Hooiveld, G. J., Dekker, J., van Mil, S. W. C., Müller, M., Kleerebezem, M., van der Meer, R., & Klaenhammer, T. R. (2015). Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(32), 10038–10043. https://doi.org/10.1073/pnas.1507645112

Jones, P. M., & George, A. M. (2004). The ABC transporter structure and mechanism: Perspectives on recent research. In *Cellular and Molecular Life Sciences* (Vol. 61, Issue 6, pp. 682–699). https://doi.org/10.1007/s00018-003-3336-9

Kamada, N., Chen, G. Y., Inohara, N., & Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. In *Nature Immunology* (Vol. 14, Issue 7, pp. 685–690). https://doi.org/10.1038/ni.2608

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. In *Nucleic Acids Research* (Vol. 28, Issue 1). http://www.genome.ad.jp/kegg/

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L., Smith, J. D., Didonato, J. A., Chen, J., Li, H., Wu, G. D., Lewis, J. D., Warrier, M., Brown, J. M., Krauss, R. M., … Hazen, S. L. (2013). Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature Medicine*, *19*(5), 576–585. https://doi.org/10.1038/nm.3145

Lesk, A. M. (2002). *Introduction to bioinformatics*. Oxford University Press.

Liu, M., Ye, Y., Jiang, J., & Yang, K. (2021). MANIEA: a microbial association network inference method based on improved Eclat association rule mining algorithm. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab241

Lynch, S. v., & Pedersen, O. (2016). The Human Intestinal Microbiome in Health and Disease. *New England Journal of Medicine*, *375*(24), 2369–2379. https://doi.org/10.1056/nejmra1600266

Maxam, A. M., & Gilbert, W. (1977). *A new method for sequencing DNA (DNA chenistry/dimethyl sulfate cleavage/hydrazine/piperidine)* (Vol. 74, Issue 2). https://www.pnas.org

Micha, R., Wallace, S. K., & Mozaffarian, D. (2010). Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: A systematic review and meta-analysis. *Circulation*, *121*(21), 2271–2283. https://doi.org/10.1161/CIRCULATIONAHA.109.924977

Nelson, D. L. (David L. (2008). *Lehninger principles of biochemistry* (M. M. Cox & A. L. Lehninger, Eds.; 5th ed. / David L...) [Book]. W. H. Freeman.

Neuman, H., Debelius, J. W., Knight, R., & Koren, O. (2015). Microbial endocrinology: The interplay between the microbiota and the endocrine system. In *FEMS Microbiology Reviews* (Vol. 39, Issue 4, pp. 509–521). Oxford University Press. https://doi.org/10.1093/femsre/fuu010

Reinhardt, C., Bergentall, M., Greiner, T. U., Schaffner, F., Östergren-Lundén, G., Petersen, L. C., Ruf, W., & Bäckhed, F. (2012). Tissue factor and PAR1 promote microbiota-induced intestinal vascular remodelling. *Nature*, *483*(7391), 627–631. https://doi.org/10.1038/nature10893

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)* (Vol. 74, Issue 12). https://www.pnas.org

Silva, Y. P., Bernardi, A., & Frozza, R. L. (2020). The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-Brain Communication. In *Frontiers in Endocrinology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fendo.2020.00025

Śliżewska, K., Markowiak-Kopeć, P., & Śliżewska, W. (2021). The role of probiotics in cancer prevention. In *Cancers* (Vol. 13, Issue 1, pp. 1–22). MDPI AG. https://doi.org/10.3390/cancers13010020

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yan, Y., Drew, D. A., Markowitz, A., Lloyd-Price, J., Abu-Ali, G., Nguyen, L. H., Tran, C., Chung, D. C., Gilpin, K. K., Meixell, D., Parziale, M., Schuck, M., Patel, Z., Richter, J. M., Kelsey, P. B., Garrett, W. S., Chan, A. T., Stadler, Z. K., & Huttenhower, C. (2020). Structure of the Mucosal and Stool

Microbiome in Lynch Syndrome. *Cell Host and Microbe*, *27*(4), 585-600.e4. https://doi.org/10.1016/j.chom.2020.03.005

Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., Nagler, C. R., Ismagilov, R. F., Mazmanian, S. K., & Hsiao, E. Y. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*, *161*(2), 264–276. https://doi.org/10.1016/j.cell.2015.02.047

Appendices

Code: https://github.com/yoyosarefun123/Alpha-Project.git

Lab report – Genome Sequencing using Illumina machines:

The Genomics Research Unit is the main genome sequencing laboratory in Tel-Aviv University. When there, we were shown and taught the four steps in genome sequencing using synthesis in Illumina machines, which are the main sequencing machines used in research nowadays.

The first step is sample prep. Sample prep is the stage where the samples are prepared, as you cannot simply enter a DNA strand and hope the sequencing machine can sequence it. The DNA samples sequenced can be either double or single strand, although additional preparations and libraries are needed to sequence single strand DNA effectively. The main part of prep is adding an adapter to the end of the DNA fragments. The adapter is a predetermined DNA strand that will be used to "bond" with the DNA strands inside the machine, that will help prime the synthesis.

The second step is cluster creation. Clusters are basically large groups of the DNA molecules entered cloned and amplified repeatedly. The reason clusters must be created is that the machine cannot measure single nucleotides inside the molecule, but when an entire cluster shows the same nucleotide sequence the machine is capable of measuring it.

The sequencing is done on a flow, which is a glass slide with lanes. Inside each lane, there's a "lawn" - small fragments of DNA that are connected to the slide and are complementary to the adapters. When the molecule "flows" over the lane, its DNA strands will connect with the lawn DNA fragment. An enzyme that synthesizes single strand DNA, also known as a polymerase, creates the complement strand of the entire DNA strand, starting from the known DNA fragment connected to the glass slide. Once the original molecule has been complemented, it's removed and we're left with a single strand DNA sequence that is the complement of the original molecule but connected to the glass slide. Finally, we start creating the actual clusters using bridge amplification. In this process, the DNA strand folds over and the adapter on the other end "bonds" with its twin adapter on the slide. The polymerase then once again synthesizes the DNA strand, and then they're split apart - and we're left with two complementary DNA strands. This process continues exponentially, until we have large clusters with the same two complementary DNA strands. Finally, all the reverse strands are cleaved and washed off, and we're left with a large amount of the same exact original DNA strand.

The third step is sequencing. The first part of this step is attaching a primer to the end of the DNA strand. The primer basically tells the polymerase where to begin synthesizing the DNA strand, and kicks off the chemical reaction. The sequencing is done per nucleotide - meaning, every single nucleotide in the molecule is sequenced separately. Basically, the polymerase synthesizes the strand up until the next nucleotide. The Illumina machine then takes a photo, in which it can separate between A, T, C, G nucleotides, and saves the image. This process is repeated until the entire strand, or the length required by the scientist, is sequenced.

The final step is data analysis - the images are taken and transformed into text files listing every single molecule, the DNA sequence of the molecule, and the grade each nucleotide has (the grade means the level of certainty that this nucleotide is correct, and accounts for margin of error).
That data is the basic data for all bioinformatic research that relates to genetic sequencing, including our own in the microbiome. In this specific research, we use pre-processed data in which different genomes and genes have already been defined and cross-referenced with KEGG database, and not the original DNA sequences like these - but our data can be directly traced to such data.