Part C:

The two habitats we chose to explore are marine and animal.

in this part we decided to filter the data in order to speed up the search and obtain more meaningful results.

We sorted the data by certain enzymes, we decided to focus on lipid transport and metabolism. we analyzed the transactions that contain cogs that relate to lipid transport and metabolism and ignored the cogs that are note related to the lipid transport and metabolism process.

Several genomes from each environment participated in the experiment:

223 genomes which contain cogs involved in the lipid transport and metabolism process.

106 of these genomes are from the marine environment and 117 of these genomes are from the animal environment.

We found out that enzymes that participant in the lipid transport and metabolism process are much more common in marine bacteria compare to animal bacteria.

For example, the most distinguishing and frequent itemset in min\_sup = 101:



The appearance of this itemset in percentages in each environment: 12.10 % in marine and 34.08 % in animal.

As we can see, enzyme that participant in lipid transport and metabolism process are much more common in marine bacteria (almost 3 times more).

We did some research and we found out about Psychrophile, extremophilic organisms that are capable of growth and reproductive in low temperature ranging from -20C to 10C. They are found in places that are permanently cold such as deep sea (temperature of seawater range from -2C to 28C). Many Psychrophile are bacteria.

The Psychrophile must also overcome the stiffening of their lipid cell membrane, as this is important for the survival and functionality of these organisms. To accomplish this, psychrophiles adapt lipid membrane structures that have a high content of short, [unsaturated fatty acids](https://en.wikipedia.org/wiki/Unsaturated_fatty_acids). Compared to longer saturated fatty acids, incorporating this type of fatty acid allows for the lipid cell membrane to have a lower melting point, which increases the fluidity of the membranes. This is the reason why Psychrophile have more enzymes that participant in lipid transport and metabolism process.

Furthermore, Psychrophile was found in whales bons 20 years after the whales died, they ate the fat that covered the whales bons.

The Psychrophile has a lot of enzymes that breakdown fat.

Most of the most distinguishing itemsets are involving cogs that are related to Acetyl Co-A activity. The main process in which acetyl Co-A participates is the synthesis and breakdown of fatty acids . Lets Take the cogs that appears in all 10 best itemsets (in min\_sup = 100):

COG0304;IQ;METABOLISM;Lipid transport and metabolism;METABOLISM;Secondary metabolites biosynthesis, transport and catabolism;3-oxoacyl-(acyl-carrier-protein) synthase;

Cog 0304 create the acyl carrier protein

COG1022;I;METABOLISM;Lipid transport and metabolism;Long-chain acyl-CoA synthetase (AMP-forming);

Cog 1022 involved in acetyl Co-A synthetase

COG0331;I;METABOLISM;Lipid transport and metabolism;Malonyl CoA-acyl carrier protein transacylase;

Cog 0331 carries the Malonyl CoA-acyl molecule.

It can be seen that enzymes involved in the process of breaking down fatty acids, and especially enzymes that are related to process’s that acetyl CoA is involved are much more common in marine environment bacteria than animal bacteria due to the explanations above.

By filtering the data by a specific process in the bacteria, the results are much better. In part B we could only run min\_sup >=200, moreover, most of the results were itemsets that contains only one cog and most of them were not distinguishing at all. Some of the itemsets was with IG = 0 and the highest IG was 0.014 (very low and not so distinguishing).

The itemset with the highest IG, appears in 103 transactions in marine and 117 (all) transactions in animal.



The information we got in part B was not informative, one reason is that the min\_sup was too high and if we tried to lower it, the running time was very high or the program failed to finish the run.

In Part C, due to the filtered data, we can lower the min\_sup and the program can finish its run. We get informative results and we can learn about the differences between the two habitats in different process in the bacteria as we described above.

By filtering the data we reduced the number of cogs that are in the data and also it decreased the size of the transactions. By that, we can lower the min\_sup and get more informative results in much faster time then part B.

In the table shown below, we compared the two algorithms. Let's pay attention that in Part C, when min\_sup is above 150, there are 106 transactions in marine and 117 in animal, so the results won't be so distinguishing.

The comparison between algorithms B and C was based on min\_sup, runtime and the quality of their results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Min\_sup** | **Running time Part B** | **Running time part C** | **Quality of result Part B** | **Quality of result Part C filtered by lipid process’s** |
| 223 | 9.446 s | 4.9 s | Itemsets that occur in all transactions in both labels – not distinguishing | There was no itemsets that appears in all transaction. |
| 220 | 15.952 s | 5 s | Itemsets that occur in most transactions in both labels – not distinguishing | There was no itemsets that appears in all transaction. |
| 210 | 10.289 s | 5 s | Itemsets that occur in most transactions in both labels – not distinguishing | There was no itemsets that appears in all transaction. |
| 200 | 21.148 s | 4.35 s | Itemsets that occur in most transaction in both labels – not distinguishing | Itemsets that occur in most transactions in both labels – not distinguishing |
| 190 | ∞ / fail | 4.34 s | No results – the program runs forever or fails to end the running | Itemsets that occur in most transactions in both labels – not distinguishing – because of high min\_sup |
| 150 | ∞ / fail | 4.55 s | No results – the program runs forever or fails to end the running | Itemsets that occur in most transactions in both labels – not distinguishing – because of high min\_sup |
| 130 | ∞ / fail | 11.207 s | No results – the program runs forever or fails to end the running | Informative results, frequent and distinguishing |
| 120 | ∞ / fail | 17.068s | No results – the program runs forever or fails to end the running | Informative results, frequent and distinguishing |
| 100 | ∞ / fail | 39.286 s | No results – the program runs forever or fails to end the running | Informative results, frequent and distinguishing |
| 95 | ∞ / fail | 6.13 s | No results – the program runs forever or fails to end the running | Informative results, frequent and distinguishing |

In conclusion, it can be seen that the algorithm in part C is faster and more importantly, return more informative results .The results show that groups of enzymes that related to the lipid transport and metabolism process’s are much more common in marine environment bacteria.

Web sites and articles we relied on in addition to our general knowledge:

1. <https://marinebio.org/oceans/temperature/> -oceans temperature
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7240044/> - Psychrophile
3. <https://www.sciencedirect.com/topics/earth-and-planetary-sciences/psychrophiles>
4. <https://ocean.si.edu/ocean-life/microbes/marine-microbes> - microbes living in the ocean
5. <https://www.livestrong.com/article/206612-enzymes-that-break-down-fat-cells/>- enzymes that break down tatt cells
6. <https://www.ncbi.nlm.nih.gov/books/NBK7919/> - fatty acids metabolism