**Part C Report**

**Maya Tanne 315342253 | Gal Bodek 316047968**

**Project Goal:** Identifying distinguishing patterns (Gene Teams) between two classes of bacteria.

**Chosen Class labels**: Animal, Plant

**Number of Animal and Plant in *bactTaxa\_habitat***: 250 (see: taxa\_of\_classes.txt)

**Number of Transactions**: 190

**Number of Animal Transactions:** 117

**Number of Plant Transactions:** 73

**Minimum Support:** 70

**Our biological additional constraints**

Our algorithm from part B takes all the transactions and finds the frequent itemsets from them, by creating subsets from the cogs.

We noticed that in *cog\_words\_bac.txt* each word (one line in the file) is attached to a specific strand, in which it was extracted from. -1 for the negative strand and 1 for the positive strand.

Our goal was to reduce the subsets that the algorithm creates by eliminating the less likely sets to be translated together to a certain biological purpose in the bacteria's cell.

Our idea was to divide the data by the two strands mentioned above. We thought that cogs that are on the same strand are more likely to be related.

**The changes we made in the algorithm**

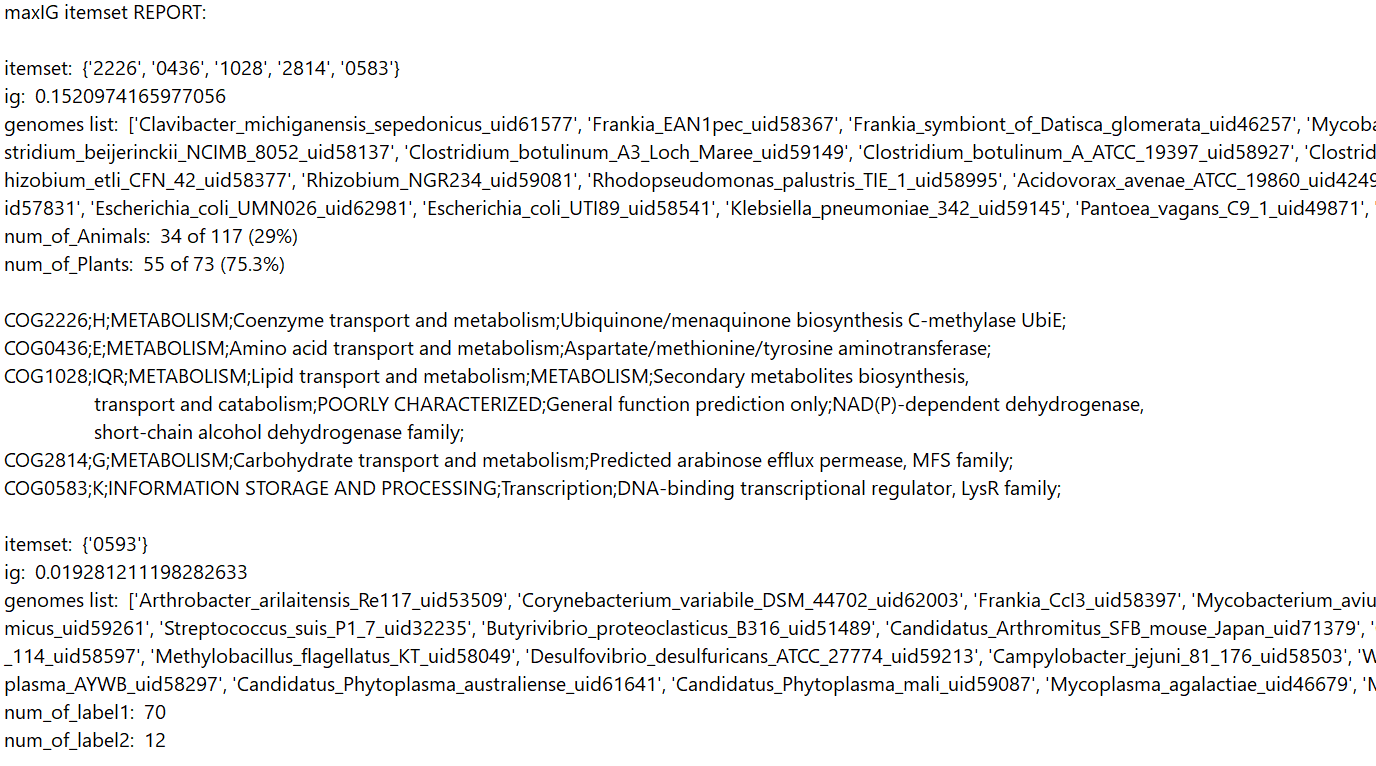
Instead of creating one FPtree for the whole data, we created two trees, one for cogs in strand 1, and one for cogs in strand -1. We ran the algorithm from part B separately on both trees (the main loop ran twice) which resulted in two separate reports. This action made the subsets that were created to have only cogs that are on the same strand and therefore, in our opinion, to be more relevant.

**The effect on the algorithm run**

The approximate run time of creating the subsets in part B was O(), where n is the number of transactions. By dividing the data into two independent trees, we reduced this run time to

O( (The number of cogs that are extracted from the -1 strand is very similar to the number of cogs that are extracted from the 1 strand).

This reduction in the run time, also enabled us to reduce the min sup to a much lower value, from 172 to 70. This made the results become much more informative.

**Report of the positive strand (1)**

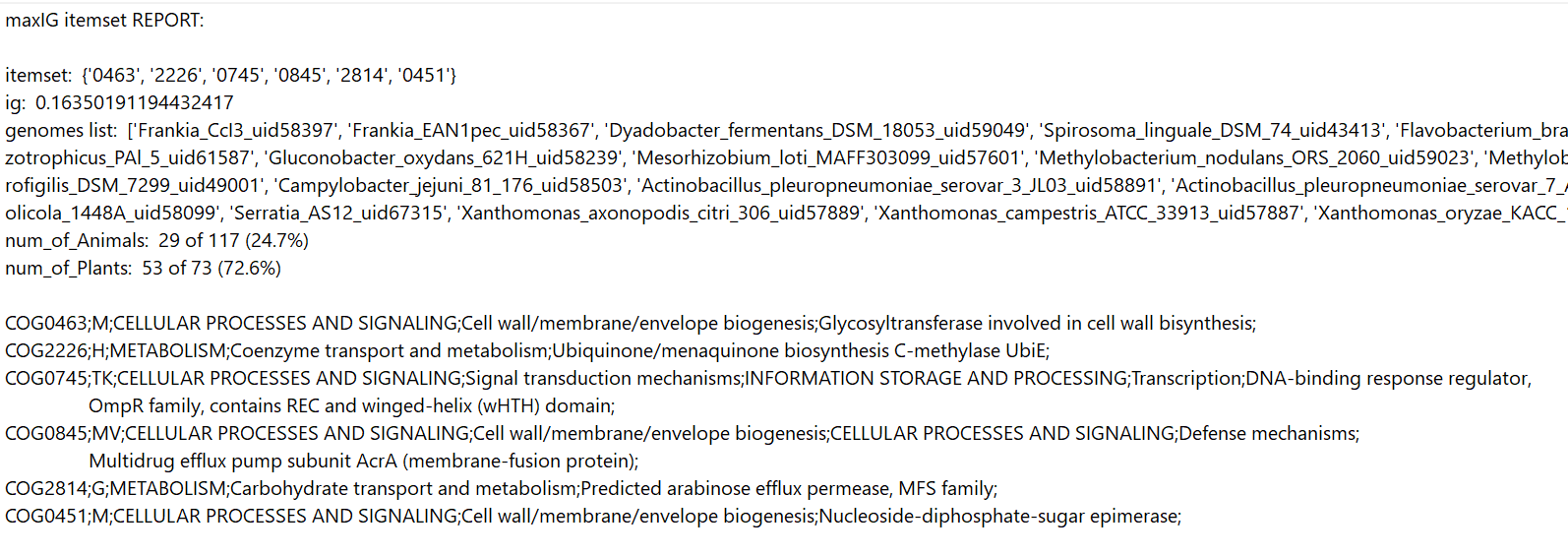
In this report we can see that the itemset which received the max IG is more likely to be in plant's bacteria then in Animals (75.3% Vs. 29%).

Let's examine the information found in biological literature about the cogs in that itemset {'2226', '0436', '1028', '2814', '0583'}.

* *COG0583* is responsible for DNA-binding transcriptional regulator which is from the **LysR family**. LysR-type transcriptional regulators (LTTRs) are the most common type of prokaryotic DNA-binding protein, and these regulatory proteins can function as either activators or repressors of gene expression.
* *COG2814* is responsible for arabinose efflux permease which is from the MFS family. **Efflux pumps** are **permeases** (membrane transport proteins) that are capable of moving a variety of different toxic compounds out of [cells](https://en.wikipedia.org/wiki/Cell_(biology)). **l-Arabinose** (l-Ara) is a plant-specific sugar accounting for 5–10% of cell wall saccharides in Arabidopsis (a family of plants).
* *COG1028* is responsible for secondary metabolites biosynthesis, transport and catabolism. **Secondary metabolites** are [organic compounds](https://en.wikipedia.org/wiki/Organic_compound) produced by [bacteria](https://en.wikipedia.org/wiki/Bacteria), [fungi](https://en.wikipedia.org/wiki/Fungi), or [plants](https://en.wikipedia.org/wiki/Plants) which are not directly involved in the normal [growth](https://en.wikipedia.org/wiki/Cell_growth), [development](https://en.wikipedia.org/wiki/Biological_development), or [reproduction](https://en.wikipedia.org/wiki/Reproduction) of the organism. Secondary metabolite synthesis in bacteria is not essential for their growth, however, they allow them to better interact with their ecological niche.
* *COG0436* is responsible for aspartate/methionine/tyrosine aminotransferase. **Aminotransferase** are a group of enzymes that catalyze the interconversion of amino acids and oxoacids by transfer of amino groups.
* *COG2226* is responsible for Ubiquinone/menaquinone biosynthesis C-methylase UbiE which is involved in respirative metabolism which responsible of producing energy to the cell.

**First report Conclusions**

From the information above, we learned that arabinose is a plant-specific sugar. We conclude that when the bacteria are based in plants their cells are dealing with a high concentration of arabinose and therefore want to control the concentration of it in the cell. They can do this by using transporters that will assist in the movement of arabinose from and into the cell (for example the arabinose efflux permease) . We assume that the bacteria cells want to activate the gene expression that will create those transporters and other proteins relevant to a metabolic pathway considering the arabinose. Activating gene expression can be done by DNA-binding transcriptional regulators. That metabolic pathway compounds may be part of the secondary metabolites of the bacteria.

**Report of the negative strand (-1)**

In this report we can see that the itemset which received the max IG is also more likely to be in plant's bacteria then in Animals (72.6% Vs. 24.7%).

Let's examine the information found in biological literature about the cogs in that itemset {'0463', '2226', '0745', '0845', '2814', '0451'}.

* *COG2814*  and *COG2226* appeared in both reports and we already examinedthem.
* *COG0745* is responsible for DNA-binding response regulator from the OmpR family, which is family of signal transduction.
* *COG0463* is responsible for Glycosyltransferase that is involved in cell wall biosynthesis. **Glycosyltransferases** are enzymes that establish natural glycosidic linkages.
* *COG0451* is responsible for Nucleoside-diphosphate-sugar epimerase. **Epimerase** are [isomerase](https://en.wikipedia.org/wiki/Isomerase) [enzymes](https://en.wikipedia.org/wiki/Enzyme) that catalyze the inversion of [stereochemistry](https://en.wikipedia.org/wiki/Stereochemistry) in biological molecules.
* *COG0845* is responsible for Multidrug efflux pump subunit AcrA (membrane-fusion protein) which is another kind of efflux pump, which we already examined.

**Second report Conclusions**

In this report and in the information above, we see similar results to the previous report. We can see that all the cogs from this itemset are translated to proteins that related to sugars. First, like in the previous report we have *COG2814* which is responsible for arabinose efflux permease and we also have another efflux pump subunit AcrA (*COG0845)*.

This strengthen our assumption that the proteins in this itemset are also related to a metabolic pathway considering the arabinose.

Glycosyltransferases establish glycosidic linkages and Nucleoside-diphosphate-sugar epimerase catalyze the inversion of [stereochemistry](https://en.wikipedia.org/wiki/Stereochemistry) of sugar molecules. We assume that their substrate in this case is arabinose. There is also another DNA-binding response regulator which, in our belief, activates the gene expression that will be translated to the proteins mentioned above.

**Summary**

As we discussed, in our belief, both itemsets from both reports are related to a metabolic pathway considering the arabinose. As we found in the biological literature, arabinose is a plant-specific sugar. From this information it is reasonable that the results we got indicate that bacteria with those genes are more likely to be found in plants.

After our biological research, we are very satisfied with the results that our algorithm produced, and we believe that they are very beneficial for Identifying distinguishing patterns between Animal and Plant classes.