Discriminative Patterns - Safari data base

Safari.py application calculates discriminative patterns between different digestion systems and different phylogenetic classes.

We calculate cogs patterns and operons patterns.

Our algorithm principles and implementation explained in the file "DiscriminativePatternsWithSupMaxPair.pptx", and in practice has two parameters which you can change (in file calcDiscriminatives.py):
- *bound* – the discriminative patterns' threshold.
- *pivot* – will separate patterns to two different files by size (bigger and smaller than the pivot).

Results:

The results files arranged in a logic and simple way:

Each compression results placed in it's unique directory, and the file name includes the *bound* and *pivot* parameter.

For example:

The directory:

/Results/**digestion**-Cogs-30/*hindgut-vs-foregut*/

Representing the comparation between two **digestion** classes, *hindgut and foregut* , based on the data file: "geneteams_30.txt", and the discriminative patterns are groups of cogs.

The directory above includes the files:

1- Discriminatives-0.8.txt,

2- Long(at_least_4)DiscriminativePatterns-0.8.txt,

3- Short(at_most_3)DiscriminativePatterns-0.8.txt.

(1) Includes all the discriminative patterns between classes *foregut* and *hindgut*, with threshold 0.8.
(2) Subgroup of (1), includes only the long discriminative patterns (patterns with at least 4 cogs).
(3) Subgroup of (1), includes only the short discriminative patterns (patterns with at most 3 cogs).

In a similar way, the directory:

\Results**phylogenetic**-Operons-40*F-vs-A*\

Representing the comparation between two **phylogenetic** classes, *F and A*, based on the data file: "OGB_catalog_plasmidome_safari_40_ins10_q1_0_q2_5_l2_instances.fasta", and the discriminative patterns are groups of operons.

Notes:

- Directory A-vs-B represents discriminative patterns that characterized class A (the leftmost class in the directory name). i.e class1 in the original algorithm is the leftmost class in the directory name.
- The classification of the animals to the specific classes based on the file "Animal_list_groups.xlsx".
- The zip includes the files: tabelsB.py and safari.py, which parsing the safari data files and launching the function "calcDisc" that we wrote in the original project.
- As we mentioned in the report, small groups comparations are less reliable. Hence, more informative phylogenetic classes comparations will include the bigger classes- A,C,F,G.