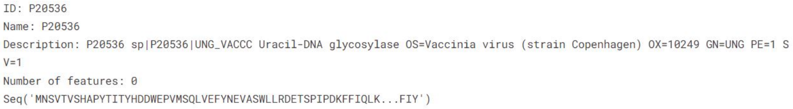
Automated Protein Function Prediction

Nishani Kasineshan

May 11, 2024

1 CAFA5-Protein Function Prediction

The goal of this project is to predict the function of proteins. The data includes following files,
train_sequences.fasta: contains aminoacid sequences for proteins in training set **ex:**
sp-P9WHI7-RECN_MYCT in the fasta header indicates the protein with UniProt ID P9WHI7 gene name RECN_MYCT taken from Swiss-Prot (sp) database.



```
ID: P20536
Name: P20536
Description: P20536 sp|P20536|UNG_VACCC Uracil-DNA glycosylase OS=Vaccinia virus (strain Copenhagen) OX=10249 GN=UNG PE=1 S
V=1
Number of features: 0
Seq('MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIPDKFFIQLK...FIY')
```

- ID: represents the unique identifier of the sequence
- Name: name associated with the sequence
- Description : additional details about the sequence
- OS : source organism
- OX : taxon id
- GN : gene name
- PE : evidence of protein existence level
- SV - sequence version
- Number of features : additional features / annotations associated with the sequence
- Seq(...) : represents the sequence data

train_terms.tsv: contains the target labels
train_taxonomy.tsv: contains the source of the organism (taxon_ids) of the protein (proteins)

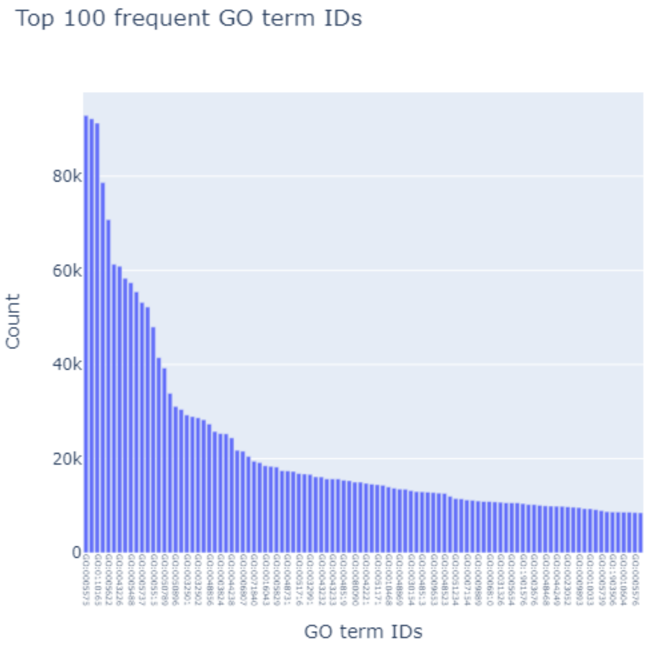
IA.txt: used for evaluation
testsuperset.fasta: contains proteins and their respective aminoacid sequences to be tested.
testsuperset_taxon_list.tsv: contains species names for taxon_ids

2 Exploratory Analysis Of GO Terms Data

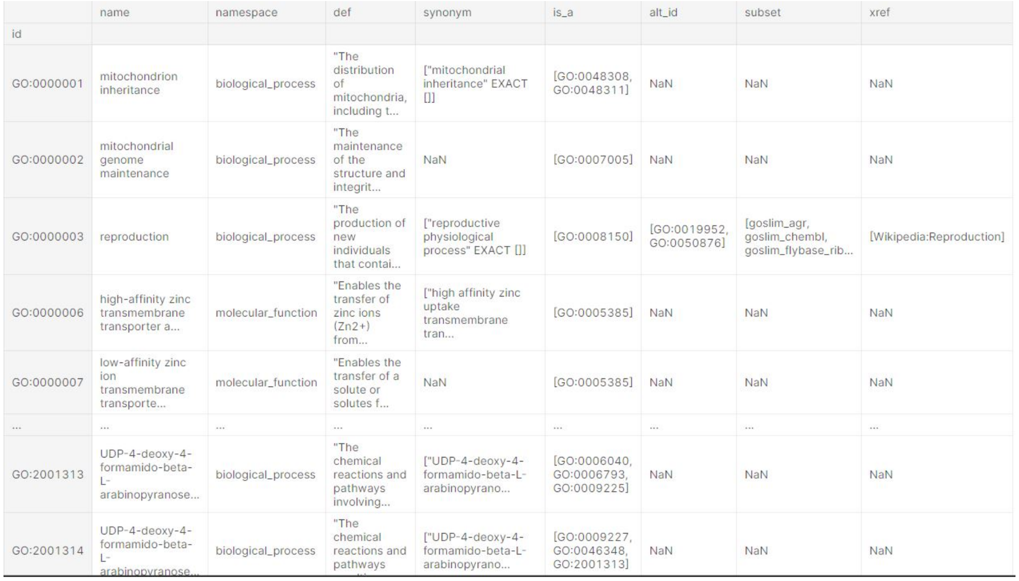
Initially an overall dataset needs to be prepared in a particular manner in order to perform analysis and further use them for models. Following steps were carried out to prepare the overall dataset:

1. The proteins ids are mapped with their taxonomyID.i.e. Dataset belonging to specific species can be extracted easily without preparing the dataset when needed.

2. The GO terms respective to the protein ids were mapped with their children, ancestors and with their union too. i.e. goatools python library provided the necessary methods to achieve this. Thus, finally the dataset will hold the dependants of each GO term respective to their protein ids.



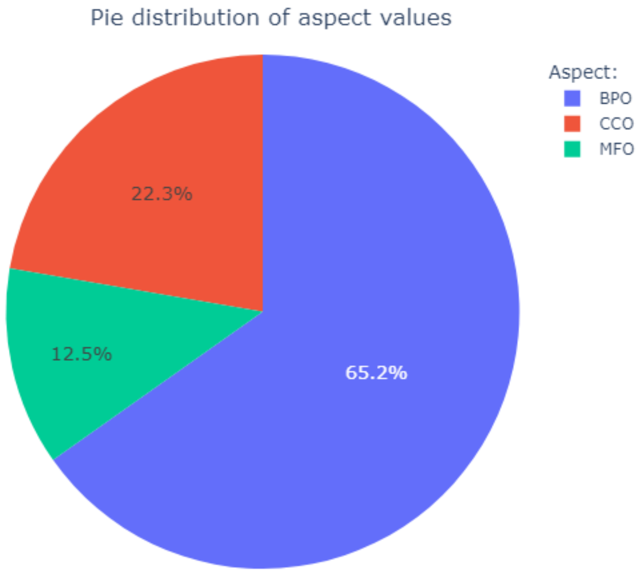
with the help of goatools, python library to inspect go file (.obo format) the whole inspection about each go term was converted into a dataframe.



	name	namespace	def	synonym	is_a	alt_id	subset	xref
id								
GO:0000001	mitochondrion inheritance	biological_process	"The distribution of mitochondria, including t...	["mitochondrial inheritance" EXACT []]	[GO:0048308, GO:0048311]	NaN	NaN	NaN
GO:0000002	mitochondrial genome maintenance	biological_process	"The maintenance of the structure and integrit...	NaN	[GO:0007005]	NaN	NaN	NaN
GO:0000003	reproduction	biological_process	"The production of new individuals that contai...	["reproductive physiological process" EXACT []]	[GO:0008150]	[GO:0019952, GO:0050876]	[goslim_agr, goslim_chembl, goslim_flybase_rib...	[Wikipedia:Reproduction]
GO:0000006	high-affinity zinc transmembrane transporter a...	molecular_function	"Enables the transfer of zinc ions (Zn2+) from...	["high affinity zinc uptake transmembrane tran...	[GO:0005385]	NaN	NaN	NaN
GO:0000007	low-affinity zinc ion transmembrane transporte...	molecular_function	"Enables the transfer of a solute or solutes f...	NaN	[GO:0005385]	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
GO:2001313	UDP-4-deoxy-4-formamido-beta-L-arabinopyranose...	biological_process	"The chemical reactions and pathways involving...	["UDP-4-deoxy-4-formamido-beta-L-arabinopyrano...	[GO:0006040, GO:0006793, GO:0009225]	NaN	NaN	NaN
GO:2001314	UDP-4-deoxy-4-formamido-beta-L-arabinopyranose...	biological_process	"The chemical reactions and pathways	["UDP-4-deoxy-4-formamido-beta-L-arabinopyrano...	[GO:0009227, GO:0046348, GO:2001313]	NaN	NaN	NaN

- name : any term have only one name defined
- def : definition
- comment : comment
- is_a : describes a subclassing relationship between one term and another
- alt_id : alternate id for this term

The following pie diagram demonstrates the distribution of the 3 GO Aspects:



2.1 Dataset Overview

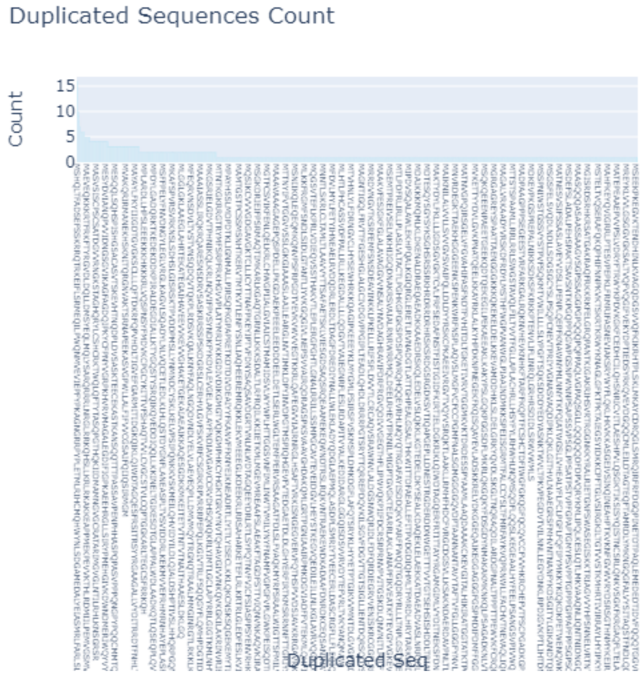
2.1.1 Analyzing train_sequences.fasta

There are 142246 unique protein ids in the training dataset. There were 3322 sequences which have duplicated sequence.

Duplicated Sequences Count
Count
15
10
5
0
Duplicated Seq

2.1.2 Analyzing testsuperset.fasta

contains protein sequences on which the participants are asked to submit predictions. There are 141864 unique protein ids in the test data with 2709 duplicated sequences.



Data	Unique
taxonomyIDs	3156
protein Ids	142246

From above table it can be seen there are 142246 unique protein ids in the dataset corresponding to 3156 unique species.

This overall dataset can be separated as 3 diffrent sub datasets based on GO term aspect: BPO,MFO,CCO.

GO Term Aspect	Unique Protein Ids
MFO	78637
BPO	92210
CCO	92912

NOTE: protein can have different function aspects related to these 3 aspects.

2.2 Specimen Data Preparation: Yeast

Yeast was taken as a specimen for data preparation where the yeast data was extracted from the overall dataset. This newly created yeast dataset contains

5484 unique protein ids corresponding to 6 types of yeast species belonging to budding yeast.

taxonomyID (budding yeast)	Yeast Species	Unique Protein Ids
559292	Saccharomyces cerevisiae S288C	5469
5478	Nakaseomyces glabratus	6
284591	Yarrowia lipolytica CLIB122	4
28985	Kluyveromyces lactis	1
284811	Eremothecium gossypii ATCC 10895	3
660122	Fusarium vanettenii 77-13-4	1

For analysis overall yeast data (5451) in train data (140588) was considered as the amount of yeast data in train only data was very low. The considered yeast data was further categorised into the three aspects where,

Data	MFO (unique)	BPO (unique)	CCO (unique)
train	78365	92066	91426
yeast	4265	4737	5201

2.2.1 Specimen Data Analysis: Yeast

Under each aspect a simple binary classification protein function prediction model and other machine learning algorithms were build for a specific GO term. Since 'protein binding' is a common function that can be found abundant in most of the proteins as follows (for yeast) GO terms with other functions which are not general were selected for analysis.

Aspect	unique protein ids
MFO	2389
BPO	-
CCO	-

The yeast data was separated to 60% train(3270) data , 20% validation(1091) data and 20% test(1090) data.

Aspect	GO term	LSTM(%)	DecisionTree(%)	SVM(%)
MFO	GO:0016787	89.3	81.9	89
BPO	GO:0034641	65.8	59.0	65.8
CCO	GO:0005622	90.7	84.1	90.7

In addition to above analysis the performance of the models were measured with different metrics.

In CAFA 5: Fmax was calculated based on weighted precision recall on each of the three test sets. Final performance was equal to arithmetic mean of the three maximum F-measures. The weight was determined by the logarithm frequency occurence of each term in large pool of terms.

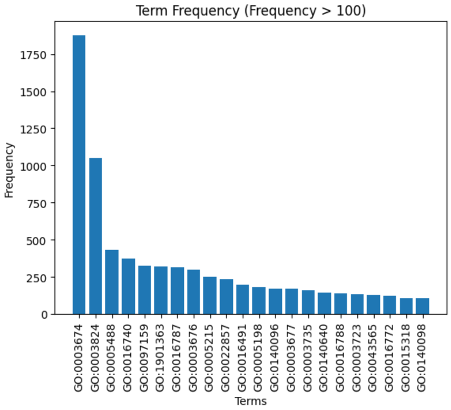
- micro average aggregates the performance across all classes as if they were a single class. Treats entire dataset as a whole. Each instance contributes equally to the overall metric regardless of its class.
- macro average ensures that each class that has the same impact on the overall metric regardless of the class imbalance.
- weighted average f1-score consider contribution of each class to the overall score based on the number of true instance in that class. classes with more instances have a greater impact on weighted average.

asp	term	acc	pr	rec	auc	f1	fmax	weighted	macro	micro
MFO	GO:0016787	89.6	0.0	0.89	52.4	18.92	20.5	17.5	11.1	10.7
MFO	GO:0016740	83.4	16.7	5.1	52.8	24.73	3.1	27.6	21.5	22.1
MFO	GO:0097159	84	12.8	4.4	49.9	25	19.94	19.7	22.1	26.6
MFO	GO:0003676	83.2	8.5	2.7	47.9	23.3	27.99	24.68	20.30	20.3
BPO	GO:0034641	65.3	36	9.9	49.5	46.87	16.2	46.4	46.4	46.4
CCO	GO:0005622	90.5	90.8	99.5	91.3	95.1	91	95.1	94.8	95.4

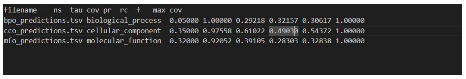
2.2.2 Average metrics

selection of GO terms for each aspect was as follows. Ignored the GO terms with relatively lower frequency because GO terms with very few positives will not have enough data to learn anything meaningful.

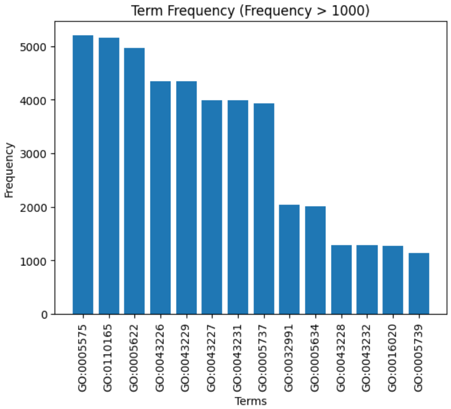
MFO GO term selection



CCO GO term selection



```
filename    ns  tau cov pr  rc  f   max_cov
bpo_predictions.tsv biological_process  0.05000 1.00000 0.29218 0.32157 0.30617 1.00000
cco_predictions.tsv cellular_component  0.35000 0.97558 0.61022 0.49030 0.54372 1.00000
mfo_predictions.tsv molecular_function  0.32000 0.92052 0.39105 0.28303 0.32838 1.00000
```



asp	term_cnt	acc	pr	rec	auc	f1	fmax	wtd	macro	micro
MFO	56(100)	92.8	6.04	4.01	50.13	12.5	15.4	14.8	22.9	24.5
BPO	78(500)	80.03	19.9	7.7	49.91	31.6		33.2		
CCO	85(100)	90.6	13.9	10.2	49.98	20.7				

2.2.3 Average metrics: Yeast

With the CAFA-Evaluator: A Python Tool for Benchmarking Ontological Classification Methods on prepared Yeast training data for each aspect following values were obtained.

Input files : Ontology file in OBO format Prediction folder containing prediction files (Tab separated file with Protein Id, Term Id and the score) Ground truth file containing targets and associated ontology terms.

Output files : A DataFrame with the evaluation results per prediction file namespace and threshold. A dictionary with the best score (max F-measure)

2.2.4 Comparison (F-max measure)

aspect	cafaeval	Ours
MFO	32.84	14.8
BPO	30.62	33.2
CCO	54.37	

2.3 Dataset Preparation

CAFA 5 Protein Function Prediction Kaggle Dataset was utilized to build the dataset. As the initial step the dataset was decided to be build as,

1. Train only dataset: represents the training data

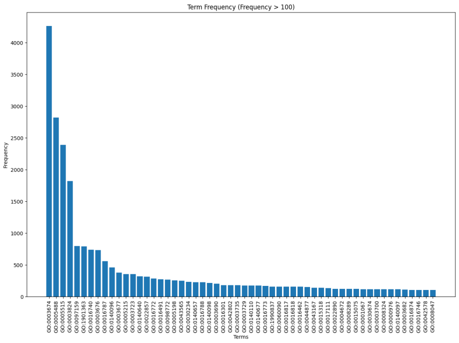


Figure 1: MFO (100)

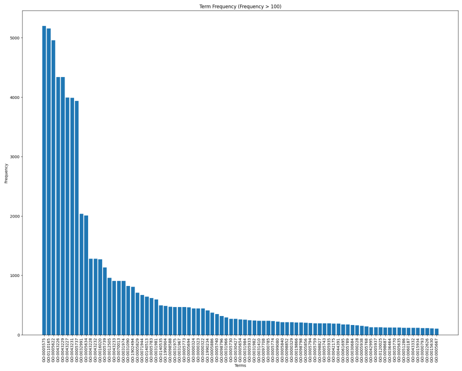


Figure 2: CCO (100)

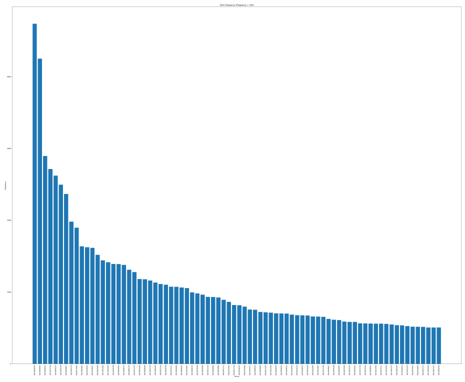


Figure 3: BPO (500)

2. Overlap dataset: represents the limited knowledge about the protein data

3. Test only dataset: represents no knowledge about the protein data

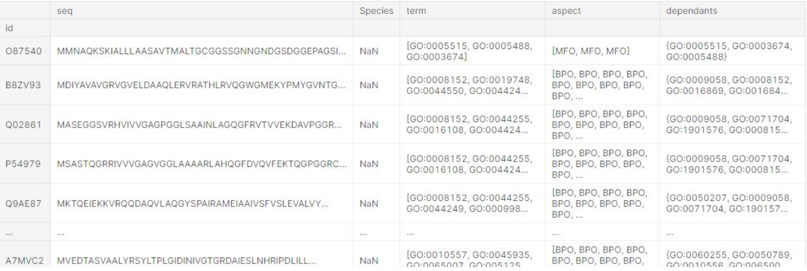
Dataset Type	Unique Protein Ids
train only	67059
overlap	73529
test only	68335

In order to prepare the train data, **train-sequences.fasta** data file was used where each protein ids (142246) are mapped to their respective protein sequences. Duplicated protein sequences found in train dataset uniquely for each species were removed and no duplicated protein ids were found. If there were duplicated protein ids, those entries have to be removed from both train and test dataset. The train data protein ids were mapped to their respective taxonomy Ids and Species names. Similarly the test data was prepared using **testsuperset.fasta** data file from the competition by removing the duplicated protein id found.

- The overlap dataset was prepared by merging the train data and the test data.
- The train only dataset was prepared by concatenating both train data and overlap data and dropped the duplicates.
- The test only dataset was prepared similarly as train only data by concatenating both test data and overlap data and dropped the duplicates.

2.4 Train Only Dataset Analysis

GO terms data from **train-terms.tsv** was mapped with the protein ids in the train data. By utilising the goatools python package a new column named dependants (term + ancestors) was created and joined with the train data.



	seq	Species	term	aspect	dependants
id					
O87540	MMNAQKSKIALLLAASAVTMALTGCGGSSGNNGNDGSDGGEPAGSI...	NaN	[GO:0005515, GO:0005488, GO:0003674]	[MFO, MFO, MFO]	{GO:0005515, GO:0003674, GO:0005488}
B8ZV93	MDIYAVAVGRVGVELDAAQLERVRATHLRVQGWGMEKYPMYGVNTG...	NaN	[GO:0008152, GO:0019748, GO:0044550, GO:004424...	[BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, ...	{GO:0009058, GO:0008152, GO:0016869, GO:001684...
Q02861	MASEGGSVRHVIVVGAGPGGLSAAINLAGQGFRVTVVEKDAVPGGR...	NaN	[GO:0008152, GO:0044255, GO:0016108, GO:004424...	[BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, ...	{GO:0009058, GO:0071704, GO:1901576, GO:000815...
P54979	MSASTQGRRIVVVGAGVGGLAAAARLAHQGFDVQVFEKTQGPGGRC...	NaN	[GO:0008152, GO:0044255, GO:0016108, GO:004424...	[BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, ...	{GO:0009058, GO:0071704, GO:1901576, GO:000815...
Q9AE87	MKTQEIEKKVRQQDAQVLAQGYSPAIRAMEIAAIVSFVSLEVALVY...	NaN	[GO:0008152, GO:0044255, GO:0044249, GO:000998...	[BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO, ...	{GO:0050207, GO:0009058, GO:0071704, GO:190157...
...	...	...	...	...	...
A7MVC2	MVEDTASVAALYRSYLTPLGIDINIVGTGRDAIESLNHRIPDLILL...	NaN	[GO:0010557, GO:0045935, GO:0065007, GO:005125	[BPO, BPO, BPO, BPO, BPO, BPO, BPO, BPO,	{GO:0060255, GO:0050789, GO:0010556, GO:006500

The train only dataset was further categorized respective to the three aspects **MFO,BPO,CCO**:

Aspect	Unique Protein Ids
MFO	26590
BPO	38547
CCO	35946

Note:yeast data is well annotated data. Since the data is dynamic same proteins may have functions in future, same data can be seen in test data too. All proteins supposed to be bind to GO terms. So GO terms related to protein binding is generally not interesting.