

main

August 23, 2021

1 GO analysis using GOATOOLS

```
[1]: import functools
import pandas as pd
import collections as cx
from pybiomart import Dataset
# GO analysis
from goatools.base import download_go_basic_obo
from goatools.base import download_ncbi_associations
from goatools.obo_parser import GODag
from goatools.anno.genetogo_reader import Gene2GoReader
from goatools.goea.go_enrichment_ns import GOEnrichmentStudyNS
```

1.1 Functions

1.1.1 Cached functions

```
[2]: @functools.lru_cache()
def get_database():
    dataset = Dataset(name="hsapiens_gene_ensembl",
                      host="http://www.ensembl.org",
                      use_cache=True)
    db = dataset.query(attributes=["ensembl_gene_id",
                                  "external_gene_name",
                                  "entrezgene_id"],
                       use_attr_names=True).dropna(subset=['entrezgene_id'])
    return db

@functools.lru_cache()
def get_deg():
    fn = '../.../_m/genes/diffExpr_EAvsAA_FDR05.txt'
    return pd.read_csv(fn, sep='\t', index_col=0)

@functools.lru_cache()
def get_ds():
    fn = "../.../visualization/_m/cluster_ds_results_annotated.txt"
```

```

return pd.read_csv(fn, sep='\t')

@functools.lru_cache()
def convert2entrez():
    df = get_deg().merge(get_ds(), left_on="Symbol", right_on="gene")
    if 'EntrezID' in df.columns:
        return df.rename(columns={'EntrezID': 'entrezgene_id'})
    else:
        return df.merge(get_database(), left_on='ensemblID',
                        right_on='ensembl_gene_id')

@functools.lru_cache()
def get_upregulated():
    df = convert2entrez()
    return df.loc[(df['t'] > 0)]

@functools.lru_cache()
def get_downregulated():
    df = convert2entrez()
    return df.loc[(df['t'] < 0)]

```

1.1.2 Simple functions

```

[3]: def obo_annotation(alpha=0.05):
    # database annotation
    fn_obo = download_go_basic_obo()
    fn_gene2go = download_ncbi_associations() # must be gunzip to work
    obodag = GODag(fn_obo) # downloads most up-to-date
    anno_hs = Gene2GoReader(fn_gene2go, taxids=[9606])
    # get associations
    ns2assoc = anno_hs.get_ns2assc()
    for nspc, id2gos in ns2assoc.items():
        print("{NS} {N:}, annotated human genes".format(NS=nspc, N=len(id2gos)))
    goeaobj = GOEnrichmentStudyNS(
        get_database()['entrezgene_id'], # List of human genes with entrez IDs
        ns2assoc, # geneid/GD associations
        obodag, # Ontologies
        propagate_counts = False,
        alpha = alpha, # default significance cut-off
        methods = ['fdr_bh'])
    return goeaobj

def run_goea(direction):

```

```

if direction == "Up":
    df = get_upregulated()
elif direction == "Down":
    df = get_downregulated()
else:
    df = convert2entrez()
geneids_study = {z[0]:z[1] for z in zip(df['entrezgene_id'], df['Symbol'])}
goeaobj = obo_annotation()
goea_results_all = goeaobj.run_study(geneids_study)
goea_results_sig = [r for r in goea_results_all if r.p_fdr_bh < 0.05]

ctr = cx.Counter([r.NS for r in goea_results_sig])
print('Significant results[{TOTAL}] = {BP} BP + {MF} MF + {CC} CC'.format(
    TOTAL=len(goea_results_sig),
    BP=ctr['BP'], # biological_process
    MF=ctr['MF'], # molecular_function
    CC=ctr['CC'])) # cellular_component

if direction == "Up":
    label = "upregulated"
elif direction == "Down":
    label = "downregulated"
else:
    label = "allDEG"
goeaobj.wr_xlsx("GO_analysis_%s.xlsx" % label, goea_results_sig)
goeaobj.wr_txt("GO_analysis_%s.txt" % label, goea_results_sig)

```

1.2 Gene ontology

```

[4]: for direction in ["All", "Up", "Down"]:
    print(direction)
    run_goea(direction)

```

All

```

requests.get(http://purl.obolibrary.org/obo/go/go-basic.obo, stream=True)
WROTE: go-basic.obo

```

```

FTP RETR ftp.ncbi.nlm.nih.gov gene/DATA gene2go.gz -> gene2go.gz
gunzip gene2go.gz

```

```

go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms

```

```

HMS:0:00:05.490595 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go

```

```

MF 18,190 annotated human genes

```

```

BP 18,505 annotated human genes

```

```

CC 19,422 annotated human genes

```

```

Load BP Gene Ontology Analysis ...

```

```

70% 20,236 of 29,107 population items found in association

```

```

Load CC Gene Ontology Analysis ...
  74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis ...
  70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 224 IDs ... 83%    160 of
192 study items found in association
  86%    192 of    224 study items found in population(29107)
Calculating 12,429 uncorrected p-values using fisher
  12,429 GO terms are associated with 17,848 of 29,107 population items
    721 GO terms are associated with    160 of    224 study items
  METHOD fdr_bh:
    0 GO terms found significant (< 0.05=alpha) (  0 enriched +  0
purified): statsmodels fdr_bh
    0 study items associated with significant GO IDs (enriched)
    0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 224 IDs ... 91%    175 of
192 study items found in association
  86%    192 of    224 study items found in population(29107)
Calculating 1,753 uncorrected p-values using fisher
  1,753 GO terms are associated with 18,711 of 29,107 population items
    223 GO terms are associated with    175 of    224 study items
  METHOD fdr_bh:
    4 GO terms found significant (< 0.05=alpha) (  4 enriched +  0
purified): statsmodels fdr_bh
    94 study items associated with significant GO IDs (enriched)
    0 study items associated with significant GO IDs (purified)

Run MF Gene Ontology Analysis: current study set of 224 IDs ... 82%    158 of
192 study items found in association
  86%    192 of    224 study items found in population(29107)
Calculating 4,420 uncorrected p-values using fisher
  4,420 GO terms are associated with 17,838 of 29,107 population items
    294 GO terms are associated with    158 of    224 study items
  METHOD fdr_bh:
    0 GO terms found significant (< 0.05=alpha) (  0 enriched +  0
purified): statsmodels fdr_bh
    0 study items associated with significant GO IDs (enriched)
    0 study items associated with significant GO IDs (purified)
Significant results[4] = 0 BP + 0 MF + 4 CC
  4 items WROTE: GO_analysis_allDEG.xlsx
  4 GOEA results for    94 study items. WROTE: GO_analysis_allDEG.txt

Up
  EXISTS: go-basic.obo
  EXISTS: gene2go

```

go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:06.068011 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go

MF 18,190 annotated human genes

BP 18,505 annotated human genes

CC 19,422 annotated human genes

Load BP Gene Ontology Analysis ...

70% 20,236 of 29,107 population items found in association

Load CC Gene Ontology Analysis ...

74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis ...

70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 119 IDs ... 80% 77 of
96 study items found in association

81% 96 of 119 study items found in population(29107)

Calculating 12,429 uncorrected p-values using fisher

12,429 GO terms are associated with 17,848 of 29,107 population items

370 GO terms are associated with 77 of 119 study items

METHOD fdr_bh:

0 GO terms found significant (< 0.05=alpha) (0 enriched + 0

purified): statsmodels fdr_bh

0 study items associated with significant GO IDs (enriched)

0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 119 IDs ... 88% 84 of
96 study items found in association

81% 96 of 119 study items found in population(29107)

Calculating 1,753 uncorrected p-values using fisher

1,753 GO terms are associated with 18,711 of 29,107 population items

130 GO terms are associated with 84 of 119 study items

METHOD fdr_bh:

0 GO terms found significant (< 0.05=alpha) (0 enriched + 0

purified): statsmodels fdr_bh

0 study items associated with significant GO IDs (enriched)

0 study items associated with significant GO IDs (purified)

Run MF Gene Ontology Analysis: current study set of 119 IDs ... 77% 74 of
96 study items found in association

81% 96 of 119 study items found in population(29107)

Calculating 4,420 uncorrected p-values using fisher

4,420 GO terms are associated with 17,838 of 29,107 population items

173 GO terms are associated with 74 of 119 study items

METHOD fdr_bh:

0 GO terms found significant (< 0.05=alpha) (0 enriched + 0

```

purified): statsmodels fdr_bh
    0 study items associated with significant GO IDs (enriched)
    0 study items associated with significant GO IDs (purified)
Significant results[0] = 0 BP + 0 MF + 0 CC
    0 items. NOT WRITING GO_analysis_upregulated.xlsx
    0 GOEA results. NOT WRITING GO_analysis_upregulated.txt
Down
    EXISTS: go-basic.obo
    EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:05.843268 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
MF 18,190 annotated human genes
BP 18,505 annotated human genes
CC 19,422 annotated human genes

Load BP Gene Ontology Analysis ...
    70% 20,236 of 29,107 population items found in association

Load CC Gene Ontology Analysis ...
    74% 21,428 of 29,107 population items found in association

Load MF Gene Ontology Analysis ...
    70% 20,354 of 29,107 population items found in association

Run BP Gene Ontology Analysis: current study set of 105 IDs ... 86%      83 of
96 study items found in association
    91%      96 of      105 study items found in population(29107)
Calculating 12,429 uncorrected p-values using fisher
    12,429 GO terms are associated with 17,848 of 29,107 population items
    417 GO terms are associated with      83 of      105 study items
    METHOD fdr_bh:
        0 GO terms found significant (< 0.05=alpha) ( 0 enriched + 0
purified): statsmodels fdr_bh
        0 study items associated with significant GO IDs (enriched)
        0 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 105 IDs ... 95%      91 of
96 study items found in association
    91%      96 of      105 study items found in population(29107)
Calculating 1,753 uncorrected p-values using fisher
    1,753 GO terms are associated with 18,711 of 29,107 population items
    161 GO terms are associated with      91 of      105 study items
    METHOD fdr_bh:
        2 GO terms found significant (< 0.05=alpha) ( 2 enriched + 0
purified): statsmodels fdr_bh
        50 study items associated with significant GO IDs (enriched)
        0 study items associated with significant GO IDs (purified)

```

```

Run MF Gene Ontology Analysis: current study set of 105 IDs ... 88%      84 of
96 study items found in association
  91%      96 of      105 study items found in population(29107)
Calculating 4,420 uncorrected p-values using fisher
  4,420 GO terms are associated with 17,838 of 29,107 population items
  167 GO terms are associated with      84 of      105 study items
METHOD fdr_bh:
  0 GO terms found significant (< 0.05=alpha) (  0 enriched +  0
purified): statsmodels fdr_bh
  0 study items associated with significant GO IDs (enriched)
  0 study items associated with significant GO IDs (purified)
Significant results[2] = 0 BP + 0 MF + 2 CC
  2 items Wrote: GO_analysis_downregulated.xlsx
  2 GOEA results for      50 study items. Wrote: GO_analysis_downregulated.txt

```

[]: