

main

August 21, 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
```

```
[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')

@functools.lru_cache()
def convert2entrez():
    df = get_deg()
    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)]

```

```

[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_ns2assoc()
    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/GO 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"
goeobj.wr_xlsx("GO_analysis_%s.xlsx" % label, goea_results_sig)
goeobj.wr_txt("GO_analysis_%s.txt" % label, goea_results_sig)

```

1.1 Gene ontology

```

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

```

```

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:0:00:04.360695 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 2967 IDs ... 80% 1,802 of
2,249 study items found in association
76% 2,249 of 2,967 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
5,001 GO terms are associated with 1,802 of 2,967 study items

```

```

METHOD fdr_bh:
    9 GO terms found significant (< 0.05=alpha) ( 7 enriched + 2
purified): statsmodels fdr_bh
    256 study items associated with significant GO IDs (enriched)
    4 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 2967 IDs ... 85% 1,917 of
2,249 study items found in association
    76% 2,249 of 2,967 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
    890 GO terms are associated with 1,917 of 2,967 study items
METHOD fdr_bh:
    37 GO terms found significant (< 0.05=alpha) ( 37 enriched + 0
purified): statsmodels fdr_bh
    1,794 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 2967 IDs ... 84% 1,881 of
2,249 study items found in association
    76% 2,249 of 2,967 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
    1,664 GO terms are associated with 1,881 of 2,967 study items
METHOD fdr_bh:
    18 GO terms found significant (< 0.05=alpha) ( 16 enriched + 2
purified): statsmodels fdr_bh
    1,609 study items associated with significant GO IDs (enriched)
    3 study items associated with significant GO IDs (purified)
Significant results[64] = 9 BP + 18 MF + 37 CC
    64 items WROTE: GO_analysis_allDEG.xlsx
    64 GOEA results for 1924 study items. WROTE: GO_analysis_allDEG.txt
EXISTS: go-basic.obo
EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.621942 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ:
gene2go
BP 18,505 annotated human genes
MF 18,190 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 1486 IDs ... 76% 809 of 1,061 study items found in association
71% 1,061 of 1,486 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
2,955 GO terms are associated with 809 of 1,486 study items
METHOD fdr_bh:
1 GO terms found significant (< 0.05=alpha) (0 enriched + 1 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 1486 IDs ... 82% 866 of 1,061 study items found in association
71% 1,061 of 1,486 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
599 GO terms are associated with 866 of 1,486 study items
METHOD fdr_bh:
7 GO terms found significant (< 0.05=alpha) (7 enriched + 0 purified): statsmodels fdr_bh
657 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 1486 IDs ... 80% 850 of 1,061 study items found in association
71% 1,061 of 1,486 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
1,015 GO terms are associated with 850 of 1,486 study items
METHOD fdr_bh:
4 GO terms found significant (< 0.05=alpha) (4 enriched + 0 purified): statsmodels fdr_bh
679 study items associated with significant GO IDs (enriched)
0 study items associated with significant GO IDs (purified)

Significant results[12] = 1 BP + 4 MF + 7 CC
12 items Wrote: GO_analysis_upregulated.xlsx
12 GOEA results for 799 study items. Wrote: GO_analysis_upregulated.txt
EXISTS: go-basic.obo
EXISTS: gene2go
go-basic.obo: fmt(1.2) rel(2021-08-18) 47,217 GO Terms
HMS:0:00:04.474433 330,313 annotations, 20,685 genes, 18,684 GOs, 1 taxids READ: gene2go
BP 18,505 annotated human genes
MF 18,190 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 1483 IDs ... 84%    994 of
1,190 study items found in association
  80% 1,190 of 1,483 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
  3,587 GO terms are associated with    994 of 1,483 study items
METHOD fdr_bh:
  8 GO terms found significant (< 0.05=alpha) ( 6 enriched + 2
purified): statsmodels fdr_bh
  76 study items associated with significant GO IDs (enriched)
  1 study items associated with significant GO IDs (purified)

Run CC Gene Ontology Analysis: current study set of 1483 IDs ... 88% 1,052 of
1,190 study items found in association
  80% 1,190 of 1,483 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
  690 GO terms are associated with 1,052 of 1,483 study items
METHOD fdr_bh:
  41 GO terms found significant (< 0.05=alpha) ( 41 enriched + 0
purified): statsmodels fdr_bh
  1,011 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 1483 IDs ... 87% 1,032 of
1,190 study items found in association
  80% 1,190 of 1,483 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
  1,153 GO terms are associated with 1,032 of 1,483 study items
METHOD fdr_bh:
  16 GO terms found significant (< 0.05=alpha) ( 14 enriched + 2
purified): statsmodels fdr_bh
  839 study items associated with significant GO IDs (enriched)
  0 study items associated with significant GO IDs (purified)
Significant results[65] = 8 BP + 16 MF + 41 CC
  65 items WROTE: GO_analysis_downregulated.xlsx
  65 GOEA results for 1065 study items. WROTE: GO_analysis_downregulated.txt

```

[]: