example_analysis

March 19, 2021

0.0.1 PySetPerm design

The pysetperm.py module includes a number of classes that provide simple building blocks for testing set enrichments. Features can be anything: genes, regulatory elements etc. as long as they have chr, start (1-based!), end(1-based) and name columns:

[29]:
```
%%bash
head -n3 data/genes.txt
```

```
chr     start   end     gene
1       904115  905037  HES4
1       921857  922761  ISG15
```

Annotations are also simply specified:

[30]:
```
%%bash
head -n3 data/kegg.txt
```

```
id      feature name
hsa00010        ACSS1   Glycolysis / Gluconeogenesis
hsa00010        ACSS2   Glycolysis / Gluconeogenesis
```

0.0.2 An example analysis

We import features and annotaions via respective classes. Features can be altered with a distance (i.e. genes +- 2000 bp). Annotations can also be filtered to have a minimum set size (i.e. at least 5 genes)

[2]:
```
import pysetperm as psp
features = psp.Features('data/genes.txt', 2000)
annotations = psp.AnnotationSet('data/kegg.txt', features.features_user_def, 5)
n_perms = 200000
cores = 10
```

Initiate test groups using the Input class:

[3]:
```
e_input = psp.Input('data/eastern_candidates.txt',
                    'data/eastern_background.txt.gz',
                    features,
                    annotations)
```

```
c_input = psp.Input('data/central_candidates.txt',
                    'data/central_background.txt.gz',
                    features,
                    annotations)

i_input = psp.Input('data/internal_candidates.txt',
                    'data/internal_background.txt.gz',
                    features,
                    annotations)
```

A Permutation class holds the permuted datasets.

[4]:
```
e_permutations = psp.Permutation(e_input, n_perms, cores)
c_permutations = psp.Permutation(c_input, n_perms, cores)
i_permutations = psp.Permutation(i_input, n_perms, cores)
```

Once permutions are completed, we determine the distribution of the Pr. X of genes belonging to Set1…n, using the SetPerPerm class. This structure enables the easy generation of joint distributions.

[7]:
```
e_per_set = psp.SetPerPerm(e_permutations,
                           annotations,
                           e_input,
                           cores)
c_per_set = psp.SetPerPerm(c_permutations,
                           annotations,
                           c_input,
                           cores)
i_per_set = psp.SetPerPerm(i_permutations,
                           annotations,
                           i_input,
                           cores)
```

Here, we can use join_objects() methods for both Imput and SetPerPerm objects, to get the joint distribution of two or more indpendent tests.

[10]:
```
# combine sims
ec_input = psp.Input.join_objects(e_input, c_input)
ec_per_set = psp.SetPerPerm.join_objects(e_per_set, c_per_set)
ei_input = psp.Input.join_objects(e_input, i_input)
ei_per_set = psp.SetPerPerm.join_objects(e_per_set, i_per_set)
ci_input = psp.Input.join_objects(c_input, i_input)
ci_per_set = psp.SetPerPerm.join_objects(c_per_set, i_per_set)
eci_input = psp.Input.join_objects(ec_input, i_input)
eci_per_set = psp.SetPerPerm.join_objects(ec_per_set, i_per_set)
```

Call the make_results_table function to generate a pandas format results table.

[11]:
```
# results
e_results = psp.make_results_table(e_input, annotations, e_per_set)
c_results = psp.make_results_table(c_input, annotations, c_per_set)
i_results = psp.make_results_table(i_input, annotations, i_per_set)
ec_results = psp.make_results_table(ec_input, annotations, ec_per_set)
ei_results = psp.make_results_table(ei_input, annotations, ei_per_set)
ci_results = psp.make_results_table(ci_input, annotations, ci_per_set)
eci_results = psp.make_results_table(eci_input, annotations, eci_per_set)
```

[33]:
```
from itables import show
from IPython.display import display
from ipywidgets import HBox, VBox
import ipywidgets as widgets
display(e_results)
```

	id	name	candidate_features	n_candidates_in_set	mean_n_resample	emp_p_e	emp_p_d	fdr_e	fdr_d	BH_fdr_e	BH_fdr_d
226	hsa04658	Th1 and Th2 cell differentiation	[CD3D, CD3G, IL12RB1, IL13, IL4, MAML3, MAPK14…	11	3.698510	0.000495	0.999880	0.092385	1.0	0.133954	0.999925
4	hsa00051	Fructose and mannose metabolism	[FUK, GMDS, HKDC1, MPI, PMM1, SORD]	6	1.466265	0.000820	0.999925	0.092385	1.0	0.133954	0.999925
47	hsa00520	Amino sugar and nucleotide sugar metabolism	[CYB5RL, FUK, GFPT2, GMDS, HKDC1, MPI, PMM1]	7	1.929355	0.001095	0.999865	0.092385	1.0	0.133954	0.999925
334	hsa05169	Epstein-Barr virus infection	[AKT3, B2M, CD3D, CD3G, HLA-A, MAPK14, NFKBIB,…	15	7.124730	0.003280	0.998825	0.164768	1.0	0.300938	0.999925
109	hsa03009	Ribosome biogenesis	[DBR1, HSPA8, MDN1, NIP7, RBM19, REXO1, REXO4,…	14	6.734525	0.004700	0.998345	0.188352	1.0	0.344978	0.999925
..	…	…	…	…	…	…	…	…	…	…	…
339	hsa05204	Chemical carcinogenesis	[]	0							
33	hsa00410	beta-Alanine metabolism	[]	0							
31	hsa00380	Tryptophan metabolism	[]	0							
349	hsa05217	Basal cell carcinoma	[]	0							
69	hsa00630	Glyoxylate and dicarboxylate metabolism	[]	0							

```
339         1.549155  1.000000  0.202404  1.000000    1.0  1.000000  0.999925
33          1.931795  1.000000  0.101269  1.000000    1.0  1.000000  0.999925
31          1.743145  1.000000  0.159219  1.000000    1.0  1.000000  0.999925
349         2.590350  1.000000  0.063135  1.000000    1.0  1.000000  0.999925
69          1.237520  1.000000  0.264189  1.000000    1.0  1.000000  0.999925

[367 rows x 11 columns]
```

[35]:
```
display(c_results)
```

```
          id                                              name  \
153  hsa04060             Cytokine-cytokine receptor interaction
369  hsa05340                               Primary immunodeficiency
317  hsa05140                                        Leishmaniasis
156  hsa04064                          NF-kappa B signaling pathway
150  hsa04050                                   Cytokine receptors
..        …                                                …
110  hsa03010                                         Ribosome
73   hsa00730                                 Thiamine metabolism
128  hsa03051                                       Proteasome
58   hsa00563  Glycosylphosphatidylinositol (GPI)-anchor bios…
74   hsa00740                               Riboflavin metabolism

                                candidate_features  n_candidates_in_set  \
153  [ACVR1, BMP6, BMP7, CCL24, CCR3, CCR9, CD4, CX…                   22
369           [ADA, AICDA, BLNK, CD4, RFX5, TNFRSF13B]                    6
317  [IFNGR2, IRAK4, ITGAM, MAPK12, MAPK13, NFKBIB,…                    8
156  [BLNK, EDARADD, ERC1, IL1R1, IRAK4, LYN, PLCG2…                   10
150  [CCR3, CCR9, CXCR6, IFNGR2, IL1R1, IL2ORA, IL3…                   10
..                                                 …                  …
110                                                []                   0
73                                                 []                   0
128                                                []                   0
58                                                 []                   0
74                                                 []                   0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e  fdr_d  BH_fdr_e  BH_fdr_d
153         6.795135  0.000005  1.000000  0.000000    1.0  0.001835       1.0
369         0.824675  0.000120  0.999990  0.010405    1.0  0.022020       1.0
317         2.049320  0.000455  0.999925  0.027655    1.0  0.055661       1.0
156         4.053805  0.004135  0.998785  0.196703    1.0  0.301142       1.0
150         3.904215  0.004380  0.998690  0.196703    1.0  0.301142       1.0
..               …         …         …         …      …         …         …
110         1.253405  1.000000  0.280364  1.000000    1.0  1.000000       1.0
73          1.111140  1.000000  0.283889  1.000000    1.0  1.000000       1.0
128         1.280525  1.000000  0.267244  1.000000    1.0  1.000000       1.0
58          0.842490  1.000000  0.412613  1.000000    1.0  1.000000       1.0
74          0.161505  1.000000  0.848071  1.000000    1.0  1.000000       1.0
```

[367 rows x 11 columns]

[34]:
```
display(i_results)
```

```
          id                                                name  \
124  hsa03036                Chromosome and associated proteins
119  hsa03021                           Transcription machinery
26   hsa00310                                Lysine degradation
113  hsa03013                                     RNA transport
374  hsa05418            Fluid shear stress and atherosclerosis
..        …                                                   …
52   hsa00534  Glycosaminoglycan biosynthesis - heparan sulfa…
152  hsa04054                      Pattern recognition receptors
51   hsa00533   Glycosaminoglycan biosynthesis - keratan sulfate
73   hsa00730                                Thiamine metabolism
55   hsa00537  Glycosylphosphatidylinositol (GPI)-anchored pr…

                                  candidate_features  n_candidates_in_set  \
124  [AHDC1, AKAP9, ALDOC, ANAPC7, ANKRD17, ARID1A,…                   84
119  [AFF1, ARID1A, ARID2, ATXN7, BRD4, CCNT1, CHD1…                   21
26   [ALDH2, ALDH3A2, ASH1L, GCDH, HADHA, KMT2D, KM…                    9
113  [AAAS, EIF2B1, EIF5B, NDC1, NUP155, NUP214, PY…                   11
374  [ACVR2A, ACVR2B, AKT2, CHUK, MAP2K5, MAPK7, NF…                   12
..                                                 …                  …
52                                                []                   0
152                                               []                   0
51                                                []                   0
73                                                []                   0
55                                                []                   0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e     fdr_d  BH_fdr_e  \
124        57.924460  0.000220  0.999865  0.040880  1.000000  0.080740
119        10.080570  0.000640  0.999785  0.058760  1.000000  0.117439
26          3.038860  0.001385  0.999820  0.087762  1.000000  0.156433
113         3.972560  0.001705  0.999535  0.087762  1.000000  0.156433
374         4.984255  0.002435  0.999180  0.092680  1.000000  0.178728
..               …         …         …         …         …         …
52          2.715120  1.000000  0.031885  1.000000  0.174461  1.000000
152         1.622385  1.000000  0.190054  1.000000  0.518005  1.000000
51          0.661145  1.000000  0.479118  1.000000  0.954243  1.000000
73          1.075285  1.000000  0.291179  1.000000  0.712622  1.000000
55          3.945295  1.000000  0.002980  1.000000  0.021366  1.000000

     BH_fdr_d
124  0.999865
119  0.999865
26   0.999865
```

```
113  0.999865
374  0.999865
..        …
52   0.486884
152  0.996872
51   0.999865
73   0.999865
55   0.099423

[367 rows x 11 columns]
```

[36]:
```
display(ec_results)
```

```
          id                                          name  \
317  hsa05140                                Leishmaniasis
226  hsa04658            Th1 and Th2 cell differentiation
153  hsa04060      Cytokine-cytokine receptor interaction
4    hsa00051             Fructose and mannose metabolism
47   hsa00520  Amino sugar and nucleotide sugar metabolism
..        …                                           …
81   hsa00830                          Retinol metabolism
82   hsa00860        Porphyrin and chlorophyll metabolism
34   hsa00430          Taurine and hypotaurine metabolism
349  hsa05217                        Basal cell carcinoma
104  hsa02042                            Bacterial toxins

                                candidate_features  n_candidates_in_set  \
317  [IL4, MAPK14, NFKBIB, PRKCB, STAT1, TAB2, IFNG…                   14
226  [CD3D, CD3G, IL12RB1, IL13, IL4, MAML3, MAPK14…                   20
153  [ACKR3, CCR9, IL12RB1, IL13, IL31, IL34, IL4, …                   31
4    [FUK, GMDS, HKDC1, MPI, PMM1, SORD, FUK, GMDS,…                   11
47   [CYB5RL, FUK, GFPT2, GMDS, HKDC1, MPI, PMM1, F…                   12
..                                                 …                   …
81                                                 []                   0
82                                                 []                   0
34                                                 []                   0
349                                                []                   0
104                                                []                   0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e     fdr_d  BH_fdr_e  \
317         4.159845  0.000005  1.000000  0.000000  1.000000  0.001835
226         7.298305  0.000010  0.999995  0.000470  1.000000  0.001835
153        13.442275  0.000015  0.999995  0.000658  1.000000  0.001835
4           2.958685  0.000030  1.000000  0.001225  1.000000  0.002752
47          4.055615  0.000170  0.999980  0.006955  1.000000  0.010748
..               …         …         …         …         …         …
81          2.617205  1.000000  0.066540  1.000000  0.774800  1.000000
82          1.553795  1.000000  0.200909  1.000000  1.000000  1.000000
```

```
34          0.890655  1.000000  0.380428  1.000000  1.000000  1.000000
349         5.069310  1.000000  0.004680  1.000000  0.321138  1.000000
104         0.084765  1.000000  0.921640  1.000000  1.000000  1.000000

     BH_fdr_d
317  1.000000
226  1.000000
153  1.000000
4    1.000000
47   1.000000
..        …
81   1.000000
82   1.000000
34   1.000000
349  0.825542
104  1.000000

[367 rows x 11 columns]
```

[37]:
```
display(ei_results)
```

```
           id                                                name  \
226  hsa04658                     Th1 and Th2 cell differentiation
26   hsa00310                                   Lysine degradation
334  hsa05169                       Epstein-Barr virus infection
124  hsa03036                  Chromosome and associated proteins
361  hsa05235  PD-L1 expression and PD-1 checkpoint pathway i…
..        …                                                   …
104  hsa02042                                    Bacterial toxins
81   hsa00830                                   Retinol metabolism
77   hsa00770                  Pantothenate and CoA biosynthesis
250  hsa04744                                    Phototransduction
82   hsa00860                Porphyrin and chlorophyll metabolism

                                    candidate_features  n_candidates_in_set  \
226  [CD3D, CD3G, IL12RB1, IL13, IL4, MAML3, MAPK14…                   19
26   [ASH1L, EHMT1, KMT2A, KMT5A, SMYD1, SMYD3, WHS…                   16
334  [AKT3, B2M, CD3D, CD3G, HLA-A, MAPK14, NFKBIB,…                   27
124  [AKAP9, ANAPC7, ANKS4B, ARHGEF10, ARID1A, ARMC…                  152
361  [AKT3, CD274, CD3D, CD3G, MAPK14, NFATC2, NFKB…                   19
..                                                 …                   …
104                                                []                   0
81                                                 []                   0
77                                                 []                   0
250                                                []                   0
82                                                 []                   0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e     fdr_d  BH_fdr_e  \
```

```
226         7.307075  0.000050  0.999975  0.009475  1.000000  0.018350
26          6.353000  0.000235  0.999945  0.023565  1.000000  0.039146
334        13.345100  0.000320  0.999900  0.023565  1.000000  0.039146
124       118.538310  0.000525  0.999650  0.027963  1.000000  0.048169
361         9.191205  0.001015  0.999660  0.043419  1.000000  0.071259
..               …       …       …       …       …       …
104         0.130180  1.000000  0.880446  1.000000  1.000000  1.000000
81          2.387655  1.000000  0.084105  1.000000  0.420211  1.000000
77          3.101705  1.000000  0.021000  1.000000  0.245089  1.000000
250         1.338225  1.000000  0.249809  1.000000  0.770701  1.000000
82          1.482415  1.000000  0.216759  1.000000  0.736863  1.000000

     BH_fdr_d
226  0.999975
26   0.999975
334  0.999975
124  0.999975
361  0.999975
..        …
104  0.999975
81   0.734914
77   0.428165
250  0.999975
82   0.999975

[367 rows x 11 columns]
```

[38]:
```
display(ci_results)
```

```
           id                                              name  \
317  hsa05140                                     Leishmaniasis
153  hsa04060           Cytokine-cytokine receptor interaction
369  hsa05340                         Primary immunodeficiency
198  hsa04350                       TGF-beta signaling pathway
226  hsa04658                 Th1 and Th2 cell differentiation
..        …                                               …
73   hsa00730                              Thiamine metabolism
51   hsa00533  Glycosaminoglycan biosynthesis - keratan sulfate
46   hsa00515                Mannose type O-glycan biosynthesis
81   hsa00830                              Retinol metabolism
116  hsa03018                                 RNA degradation

                                  candidate_features  n_candidates_in_set  \
317  [IFNGR2, IRAK4, ITGAM, MAPK12, MAPK13, NFKBIB,…                   13
153  [ACVR1, BMP6, BMP7, CCL24, CCR3, CCR9, CD4, CX…                   29
369  [ADA, AICDA, BLNK, CD4, RFX5, TNFRSF13B, BLNK,…                    8
198  [ACVR1, BMP6, BMP7, FBN1, GREM1, INHBA, LTBP1,…                   21
226  [CD4, DLL1, IFNGR2, IL13, IL4R, MAPK12, MAPK13…                   17
```

```
..                                                    …                   …
73                                                    []                  0
51                                                    []                  0
46                                                    []                  0
81                                                    []                  0
116                                                   []                  0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e     fdr_d  BH_fdr_e  \
317         4.095365  0.000075  0.999985  0.014505  1.000000  0.014068
153        13.523935  0.000110  0.999945  0.014505  1.000000  0.014068
369         1.576295  0.000115  0.999980  0.014505  1.000000  0.014068
198         9.623055  0.000200  0.999940  0.014505  1.000000  0.018350
226         7.208360  0.000565  0.999830  0.024924  1.000000  0.041471
..               …       …         …         …         …         …
73          2.186425  1.000000  0.083050  1.000000  0.554494  1.000000
51          1.286265  1.000000  0.242634  1.000000  0.941195  1.000000
46          2.905925  1.000000  0.029310  1.000000  0.357775  1.000000
81          2.518320  1.000000  0.074920  1.000000  0.532925  1.000000
116         4.608260  1.000000  0.008305  1.000000  0.132108  1.000000

     BH_fdr_d
317  0.999985
153  0.999985
369  0.999985
198  0.999985
226  0.999985
..        …
73   0.999985
51   0.999985
46   0.705176
81   0.948121
116  0.319839

[367 rows x 11 columns]
```

[39]:
```
display(eci_results)
```

```
          id                                    name  \
226  hsa04658        Th1 and Th2 cell differentiation
317  hsa05140                            Leishmaniasis
334  hsa05169            Epstein-Barr virus infection
369  hsa05340                Primary immunodeficiency
153  hsa04060  Cytokine-cytokine receptor interaction
..        …                                       …
75   hsa00750                   Vitamin B6 metabolism
104  hsa02042                        Bacterial toxins
118  hsa03020                          RNA polymerase
74   hsa00740                   Riboflavin metabolism
```

```
82   hsa00860    Porphyrin and chlorophyll metabolism

                                  candidate_features  n_candidates_in_set  \
226  [CD3D, CD3G, IL12RB1, IL13, IL4, MAML3, MAPK14…                   28
317  [IL4, MAPK14, NFKBIB, PRKCB, STAT1, TAB2, IFNG…                   19
334  [AKT3, B2M, CD3D, CD3G, HLA-A, MAPK14, NFKBIB,…                   39
369  [CD3D, TNFRSF13B, ADA, AICDA, BLNK, CD4, RFX5,…                   10
153  [ACKR3, CCR9, IL12RB1, IL13, IL31, IL34, IL4, …                   38
..                                                 …                  …
75                                                []                   0
104                                               []                   0
118                                               []                   0
74                                                []                   0
82                                                []                   0

     mean_n_resample   emp_p_e   emp_p_d     fdr_e     fdr_d  BH_fdr_e  \
226        10.906870  0.000010  1.000000  0.000545  1.000000  0.001835
317         6.205890  0.000010  1.000000  0.000545  1.000000  0.001835
334        19.527630  0.000015  1.000000  0.000687  1.000000  0.001835
369         2.458895  0.000130  0.999970  0.005809  1.000000  0.009542
153        20.171075  0.000130  0.999945  0.005809  1.000000  0.009542
..               …         …         …         …         …         …
75          0.612195  1.000000  0.530487  1.000000  1.000000  1.000000
104         0.154295  1.000000  0.861841  1.000000  1.000000  1.000000
118         2.140525  1.000000  0.109329  1.000000  0.623935  1.000000
74          0.555795  1.000000  0.564242  1.000000  1.000000  1.000000
82          2.239385  1.000000  0.098825  1.000000  0.609766  1.000000

     BH_fdr_d
226  1.000000
317  1.000000
334  1.000000
369  1.000000
153  1.000000
..        …
75   1.000000
104  1.000000
118  0.866453
74   1.000000
82   0.866453

[367 rows x 11 columns]
```

[ ]: