*-00-hit
1
*-01-hit-cog
3
*-01b-hit-cog-cat
0
01b-hit-cog-cat
cog_abundance
2
*-02a-hit-cog-mpp
4
*-02b-hit-cog-mpp-cop
3
*-03a-hit-cog-mpm-cop-nul
6
*-03b-hit-cog-mpm-cop-nul-nve
7
*-03c-hit-cog-mpm-cop-nul-nve-nve
9
*-04a-hit-cog-mpm-cop-nul-nve-nve-xpe
0
04a-hit-cog-mpm-cop-nul-nve-nve-xpe
module_coverage
8
*-04b-hit-cog-mpm-cop-nul-nve-nve
0
04b-hit-cog-mpm-cop-nul-nve-nve
module_abundance
5
*-03a-hit-cog-mpm-cop-wbl
6
*-03b-hit-cog-mpm-cop-wbl-nve
7
*-03c-hit-cog-mpm-cop-wbl-nve-nve
9
*-04a-hit-cog-mpm-cop-wbl-nve-nve-xpe
0
04a-hit-cog-mpm-cop-wbl-nve-nve-xpe
module_coverage_with_wb_smoothing
8
*-04b-hit-cog-mpm-cop-wbl-nve-nve
0
04b-hit-cog-mpm-cop-wbl-nve-nve
module_abundance_with_wb_smoothing
2
*-02a-hit-cog-mpt
4
*-02b-hit-cog-mpt-cop
3
*-03a-hit-cog-mpt-cop-nul
6
*-03b-hit-cog-mpt-cop-nul-nve
7
*-03c-hit-cog-mpt-cop-nul-nve-nve
9
*-04a-hit-cog-mpt-cop-nul-nve-nve-xpe
0
04a-hit-cog-mpt-cop-nul-nve-nve-xpe
pathway_coverage
8
*-04b-hit-cog-mpt-cop-nul-nve-nve
0
04b-hit-cog-mpt-cop-nul-nve-nve
pathway_abundance
5
*-03a-hit-cog-mpt-cop-wbl
6
*-03b-hit-cog-mpt-cop-wbl-nve
7
*-03c-hit-cog-mpt-cop-wbl-nve-nve
9
*-04a-hit-cog-mpt-cop-wbl-nve-nve-xpe
0
04a-hit-cog-mpt-cop-wbl-nve-nve-xpe
pathway_coverage_with_wb_smoothing
8
*-04b-hit-cog-mpt-cop-wbl-nve-nve
0
04b-hit-cog-mpt-cop-wbl-nve-nve
pathway_adundance_with_wb_smoothing

1 *hits2enzymes.py*: generation fo COG abundance with

G_{cog_i} set of genes related to cog_i
R_g set of reads similar (given similarity search) to gene g
G_r set of genes similar (given similarity search) to read r

$$abundance_1(cog_i) = \sum_{g \in G_{cog_i}} \frac{1}{length(g)} \sum_{r \in R_g} \frac{\exp(-evalue(r))}{max\left(10^{20}, \sum_{g' \in G_r} \exp(-evalue(g'))\right)}$$

2 *enzymes2pathways_mp.py*: association between COG and module (mpp) or pathways (mpt)

3 *cat.py*: elimination of empty and comment lines

4 *taxlim.py*: filtering by taxonomic limitation and copy number normalization (option) with

O_{cog_i} set of organisms found for cog_i
$G_{O|cog_i}$ set of genes for cog_i for the organism o

$$abundance_2(cog_i) = abundance_1(cog_i) \frac{\sum_{g \in G_{O|cog_i}} abundance(g|o)}{\sum_{o \in O_{cog_i}} abundance(o)}$$

5 *smooth_wb.py*: abundance smoothing using Witten-Bell discounting (reevalution of 0 and small probabilities) with

iT number of COG found

$$dN = \sum_{\text{cog found}} \sum_{m/p} \frac{abundance(m/p|cog)}{number(m/p)}$$

$$abundance_3(cog_i) = \begin{cases} abundance_2(cog_i) \frac{dN}{dN+iT} & \text{if } cog_i \text{ is found in pathway/module} \\ dN \frac{iT}{number(\text{cog not found})} & \text{else} \\ 0 & ? \end{cases}$$

6 *gapfill.py*: increase of the effective contribution of unobserved members of otherwise abundant pathways within each retained pathway/modules. COG with relative abundance 1.5 interquartile ranges below the pathway median are boosted to an effective abundance equal to median for purposes of subsequent calculation

7 *pathcov.py*: coverage calculation to indicate the likelihood that all genes needed to operate the pathway/module are present

pathway coverage: fraction of COG in the pathway that were confidently present, specifically with abundance greater than the overall sample median

$$cov_p = \frac{1}{|p|} \sum_{i \in p} \partial \left(\omega_{i,p} > \widetilde{\omega}_{i,p}\right)$$

module coverage: harmonic mean of the X^2 CDF with $\widetilde{\omega}_{i,p}$ degrees of freedom evaluated at each $\omega_{i,p}$ for each required $i \in m$, maximizing over optional genes *i* and alternative submodules

8 *pathab.py*: abundance calculation for each pathway/module

pathway abundance

$$abd_p = \frac{2}{|p|} \sum_{i \in [p/2]} \omega_{i,p}$$

module abundance: harmonic mean of the sample gene family abundances

9 *pathcov_xp.py*: eliminate pathways/modules with low coverage

0 *merge_table.py*: addition of category names
zero.py: remove extra-spaces and weird values
filter.py: remove pathways/modules with less than 4 COG
normalize.py: normalize abundance/coverage values to have a sum equal to 1
eco.py: ecological statistic calculation (inverse Simpson index, Shannon richness index, Pielou's eveness index, richness)