

taxon-utils之翻译指定分类水平主要分类的世系：translate

一、taxon-utils translate介绍

功能描述：

`taxon-utils translate` 指定的特定节点或者节点列表，使用构建的节点树文件，回溯到根节点，输出中间节点的路径（界，门，纲，目，科，属，种， 示例

`k:Bacteria,p:Firmicutes_A,c:Clostridia,o:Lachnospirales,f:Lachnospiraceae,g:14-2,s:14-2 sp001940225`）以及提供对应的数字等级路径（`3,64,222,873,2385,3766,11673`），适用于微生物组数据分析。

命令行接口：

```
1 $ taxon-utils translate
2
3 Usage: taxon-utils translate [options] <taxon.map> <tab|taxonIds>
4
5 Options:
6   -c INT    column for taxon translate, default: [2]
7   -n        print lineage for specified node, using ',' for multi taxonId.
```

可选参数：

1	-c	整型	指定转化的列，默认为第二列；
2	-n		输出指定节点的世系

二、使用场景实例及其用法

示例演示：

示例文件在: `data` 目录

示例文件: `classify.txt.gz`, `taxon.map.gz`

```
1 $ zcat classify.txt.gz | head -n 6
```

```

1 C      A01050:204:HF7FGDSXY:4:1101:12943:1016  25457  150|149 0:22 25457:4
0:7 25458:5 0:78 |:| 3:4 3343:5 8109:2 25457:5 25460:1 25457:6 25693:5 0:48
25458:2 0:1 3343:2 14694:7 3343:10 3:5 0:12
2 C      A01050:204:HF7FGDSXY:4:1101:12337:1031  18331  150|149 0:60 31294:5
0:25 18331:3 0:23 |:| 0:93 18331:3 0:19
3 C      A01050:204:HF7FGDSXY:4:1101:16866:1047  8364  150|150 0:17 20367:2
35825:1 0:8 3:4 20616:5 3:1 4523:5 3:3 0:9 3:5 0:19 27173:2 0:12 27173:2 0:21
|:| 0:2 2878:3 24660:5 28576:3 30576:2 3:13 0:9 2483:17 27170:3 2483:2
27170:5 3:6 8364:3 3:7 8364:3 27172:2 8364:5 27172:6 171:5 0:15
4 C      A01050:204:HF7FGDSXY:4:1101:22742:1047  23158  150|150 0:82 23158:4
0:30 |:| 0:116
5 C      A01050:204:HF7FGDSXY:4:1101:28664:1063  559  150|150 0:22 559:3
0:7 171:5 0:79 |:| 0:116
6 C      A01050:204:HF7FGDSXY:4:1101:20473:1094  25986  150|150 0:45 25986:3
0:68 |:| 0:116

```

注意事项: `classify` 文件为 `Kraken2` 分类的结果, 第一列为标识符: `C`: 可以分类的序列, `U`: 不能分类的序列, 第三列为分类的 `Taxonomy ID`, 可以为 `NCBI` 分类号或者使用 `GTDB` 的自定义分类号, 正常我们只需要第二列和第三列。

```
1 $ zcat taxon.map.gz | head -n 6
```

```

1 1      1      no rank root    root
2 2      1      superkingdom  Archaea root
3 3      1      superkingdom  Bacteria      root
4 4      3      phylum  4572-55 Bacteria
5 5      3      phylum  AABM5-125-24 Bacteria
6 6      3      phylum  AB1-6   Bacteria

```

运行命令:

```
1 $ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6
```

```

1 A01050:204:HF7FGDSXY:4:1101:12943:1016  25457
2 A01050:204:HF7FGDSXY:4:1101:12337:1031  18331
3 A01050:204:HF7FGDSXY:4:1101:16866:1047  8364
4 A01050:204:HF7FGDSXY:4:1101:22742:1047  23158
5 A01050:204:HF7FGDSXY:4:1101:28664:1063  559
6 A01050:204:HF7FGDSXY:4:1101:20473:1094  25986

```

将第二例翻译成物种世系格式:

```
1 $ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6 | taxon-utils
  translate -c 2 taxon.map.gz -
```

```

1 | A01050:204:HF7FGDSXY:4:1101:12943:1016 25457
   | k:Bacteria,p:Eremiobacterota,c:Eremiobacteria,o:UBP12,f:UBA5184,g:Palsa-
   | 1478,s:Palsa-1478 sp003140215 3,56,265,1390,3343,8109,25457
2 | A01050:204:HF7FGDSXY:4:1101:12337:1031 18331
   | k:Bacteria,p:Gemmatimonadota,c:Gemmatimonadetes,o:Gemmatimonadales,f:Gemmatim
   | onadaceae,g:Fen-1231,s:Fen-1231 sp003171215
   | 3,78,283,806,2237,5826,18331
3 | A01050:204:HF7FGDSXY:4:1101:16866:1047 8364
   | k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Micrococca
   | ceae,g:Pseudarthrobacter 3,9,171,559,2483,8364
4 | A01050:204:HF7FGDSXY:4:1101:22742:1047 23158
   | k:Bacteria,p:Proteobacteria,c:Alphaproteobacteria,o:Rhizobiales,f:Beijerincki
   | aceae,g:Methylobacterium,s:Methylobacterium sp003173775
   | 3,110,175,1021,1696,7507,23158
5 | A01050:204:HF7FGDSXY:4:1101:28664:1063 559
   | k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales
   | 3,9,171,559
6 | A01050:204:HF7FGDSXY:4:1101:20473:1094 25986
   | k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Dermatophi
   | laceae,g:Pedococcus,s:Pedococcus sp001426245
   | 3,9,171,559,1919,8203,25986

```

多输出两列，一列为主要分类水平: **k,p,c,o,f,g,s** 界门纲目科属种分类， 一列为对应的数值编号， 比如：

k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Dermatophilaceae,g:Pedococcus,s:Pedococcus sp001426245，分类之间使用 **,** 分割。

命令行参数 **1: -n** 指定需要翻译的数字编号， 比如: **25986**。

```

1 | $ taxon-utils translate -n taxon.map.gz 25986

```

```

1 | k:Bacteria,p:Actinobacteriota,c:Actinobacteria,o:Actinomycetales,f:Dermatophi
   | laceae,g:Pedococcus,s:Pedococcus sp001426245
   | 3,9,171,559,1919,8203,25986

```

本文材料为 **BASE (Biostack Applied bioinformatic SEies)** 课程 **Linux Command Line Tools for Life Scientists** 材料， 版权归 上海逻捷信息科技有限公司 所有。

Last Update: 2020-08-10 11:56 AM