Examples of Regular Expressions

Here are some example of typical and useful regular expressions.

All the working code is also in this folder on github or you can check out the repository. https://github.com/biodataprog/GEN220_2019_examples/tree/master/Python_5

Data/text example 1: FASTA header information

Often FASTA files will be used to encode a lot of information and all of this is stored in the header line. We want to get the following out of this header into some variables so we can do something more with the data.

- Species prefix (`Hsapiens`)
- Accession number (`ABC10021.1`)
- GI number (`1133455`)
- Gene name (`YFG1`)
- Description string (`This gene makes and enzyme`)

The data look like this and are stored in a file called `sequences.fas`

```
>Hsapiens|ABC10021.1|gi|1133455 GENE=YFG1 DESC="This gene makes an enzyme"
ATGAATGGACAGAGTA
```

```
#!/usr/bin/env python3
import re

pat = re.compile(r'>([^\|]+)\|([^\|]+)\|gi\|(\d+)\s+GENE=(\S+)\s+DESC=\"([^\"]+)\"')
with open("sequences.fas","r") as fh:
    for line in fh:
        if line.startswith(">"):
            m = pat.match(line)
            print(line.strip())
            if m:
                species=m.group(1)
                acc    =m.group(2)
                gi     =m.group(3)
                gene   =m.group(4)
                Desc   =m.group(5)
                print("species   = ",species)
                print("accession = ",acc)
                print("gi.       = ",gi)
                print("Gene name = ",gene)
                print("Desc is.  = '{}'".format(Desc))
```

Simple delimiting

“‘python text=[“ABC10..30”, “ABC 30..40”]

for r in text: