

Practical 5. String and file operation

Chaochen Wang

IBI1, 2019/20

Learning Objects

- Create and manipulate strings in Python
- Search and find regular expression patterns in strings in Python
- Read and write files in Python

You have learned how to use Github in previous practical. **Please create a directory Practical 8 in your Github repository IBI1 2019-20 and put all your Python scripts in this practical there.**

1. Reverse complementary sequence

In biological studies, we face a lot of sequence data (DNA, RNA and proteins). Under certain circumstances, we need to know the reverse complementary sequence of a DNA sequence. For example, the reverse complementary sequence of 'ATGCAC' is 'GTGCAT'.

Please create a new Python script file **RC.py** in Spyder and create a string variable seq.

```
seq = 'ATGCGACTACGATCGAGGGCCAT'
```

then write codes to print out the reverse complementary sequence of it.

2. Get mitochondria gene sequence in fasta format

A common format to store the sequence data is FASTA. In a fasta/fa file, each sequence starts with a ">" symbol followed by the name of the sequence and the description of the sequence and the remaining line(s) contain the sequence itself. Here is a fasta file downloaded from ftp://ftp.ensemblgenomes.org/pub/fungi/release-46/fasta/saccharomyces_cerevisiae/cdna/. The file contains the cDNA sequences of all the genes in yeast *Saccharomyces Cerevisiae*. You can view the first lines of the file by using **head** command in your terminal (PC users can use Cypwin) as shown here. The name line has a list of gene information, such as the position in the chromosome, the gene name, the biotype, etc.,

This gene is on chromosome II

```

$ head -20 Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa
>YBR024W_mRNA cdna chromosome:R64-1-1:II:289445:290350:1 gene:YBR024W gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:SCO2 description:Protein anchored to mitochondrial inner membrane; may have a redundant function with Sco1p in delivery of copper to cytochrome c oxidase; interacts with Cox2p; SCO2 has a paralog, SCO1, that arose from the whole genome duplication [Source:SGD;Acc:S000000228]
ATGTTGAATAGTTCAAGAAAAATATGCTTGTGCTCCCTATTTCAGACAAGCGAACGCTCTCA
ATAAAAGGACTCTTTTATAATGGAGGCGCATATCGAAGAGGGTTTCAACGGGATGTTGT
TTGAGGAGTGATAACAAGGAAAGCCCAAGTGAAGACAACCACTAGATAGGCTACAACCTA
GGTGATGAAATCAATGAACCAAGAGCTATTAGAACCAGGTTTTTCAATTTCCAGATGG
AAGGCCACCATGCTCTATTGTTGCTAAGTGGTGGGACGTATGCCTATTTATCAAGAAAA
AGACGCTTGTAGAACTGAAAAGGAAGCAGATGCTAACAGAGCTTACGGTTCAGTAGCA
CTTGGCGGTCCTTTCAATTTAACAGATTTTAAATGGTAAGCCTTTCACTGAGGAGAATTTG
AAGGGTAAGTTTTCAATTTTATCTTTGGATTTCAGTCATTGCCCGGACATTTGTCCAGAA
GAGCTTGACAGATTAAAGTATTGGATTCTGAATTAGATGATAAAGACCATATAAAGATA
CAGCCATTGTTTATCTCATGTGATCCTGCAAGAGATACACCGGATGCTTTGAAAGAGTAC
TTAAGCGATTTTCAACCGATATCAATGTTTAAACGGTACGTACGACCAAGTGAAGAAC
GTATGCAAAAAATACAAGGTATATTTTCAACTCCACGTGATGTCAAGCCCAACAGGAT
TACTTAGTGGACCATTCGATATTTTCTATTTGATCGACCTGAAGGACAGTTTATCGAT
GCGTTGGGAAGAACTACGATGAGCAATCTGGTCTCGAAAAGATTCTGTGAACAAATTCAG
GCGTATGTGCCAAAGGAAGAACGGGAGCGTAGGTCAAAAAATGGTACTCTTTTATCTTC
AATTGA
>YDL245C_mRNA cdna chromosome:R64-1-1:IV:11657:13360:1 gene:YDL245C gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:HXK15 description:Putative transmembrane polyol transporter; supports growth on and uptake of mannitol, sorbitol and xylitol with moderate affinity when overexpressed in a strain deleted for hexose family members; minor hexose transport activity when overexpressed in a similar strain; similarity to hexose transporters; expression is induced by low levels of glucose and repressed by high levels of glucose [Source:SGD;Acc:S000002404]
ATGGCAAGCGAAGCAGTCCTCACCAGAAATTAATGCAGATAATCTAAACAGTAGTGCAGCT
GACGTTTCATGTACAGCCACCCGAGAGAGAAATGGTCAGACCGGTTTTATGACAAAGAA

```

The sequence information in one line starts with '>'

Gene name

Sequence in several lines

This gene is on chromosome IV

Download the file `Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa` from Learn. Starting from this file, now we want to extract all the genes on the mitochondria chromosome (There are 17 chromosomes in the species, named as I, II, III...XVI and Mito, among which the 'Mito' is the mitochondria chromosome). Create a new script file `get_mito_gene.py`. In it, write codes to read the file, extract the sequences of mitochondria genes and simplify the sequence name by only putting the gene name and the length of the sequence. Output the result in a new fasta file `mito_gene.fa`.

3. Reverse complementary sequences of mitochondria genes

Now let us combine your techniques above!

We are looking for a python script '`Mito_RC.py`' (use Spyder to write):

- 1) asks for the user to input a filename as the new fasta file;
- 2) stores the reverse complementary sequences of mitochondria genes in fasta format (do not put line breaks in the sequence as in the original file, so that the entire sequence is in one line);
- 3) the sequence name of each gene only has the gene name and sequence length.

The first lines in the file should be similar to this:

```

$ head -10 rc.fa
>Q0045 1605
TTAAGATGGTACAGCTGGGTGATTAATAGAGGTACAGCTGGTGGAGAAGTTAATAAGAAATCGATAGATGAAGATTTAACTGTATTTAAATTAAGATAGGATTAGATTCTACAAATCAGGTGCTTTATTATAAATACTGATTTATTATTAACCTTTATTGTTTAACTCA
TTAACTAATTGATCATATAAATATAGATAAATAAGAAATAATGATAATAGTGAATGAATGAACCAATAGAAGCGACATAATTTTCATCCTGCGAAAGCATCAGGATAATCAGGAATTTCTTAGGGCATACCATTAAATACCTAAAAATGCATTGGGAAGAAAAATCACTAG
CCCCAATGAATAATATCAGAAATTTGAATTTGAGCTAATTTTTTCATTATAGTTTAAACCTAAAAATTTTGAGGACTTCAATAATAGTATCCTGCAAAATAAGAGAAAAATAGCACCCATTGATAATACATAGTGAAAAATGTCCACCACGTAGTAAGTATCGTGGAAATGCTACATC
TAATGAGGGCGTTAGCTAAGGCAACACCAGTTAAACCACCCATTGTGAATAAGAAATAAGAAATGCAATTGCATATAACATAGGTAGTGCTAACTCTAATTGAACCACCATGGATTAGAGCTAATCATGAGAAAAATTTTAATTCCTGTGGAAATGCAATATCATTAGTGAGATC
AGGAAATATGCTCTAAGATCTGCATCTAATCCTACAATATACATATGATGTGATCATACTAAGAATCCTAATAATCCAATTGAAGCCATAGCATATACCATTGAAATTTACCAAAATACAGGTTTTTTAGAATATGTTGATACACTGATGTGAAATAATACCAATCCAGGAA
TAATTAATAATATATCTTCAGGGTGACCAAGAATCAAAATAAATGCTCGTATAAGATTGGGTCAACCACTCCTGATACCTCAAGAATAAGAAATTTGAAGTTTCTATCTAATAATAACATTGTAATACCAGCAGATAATACAGGTAAATGATAATAATAAAGACGCTGT
AATGAAATGATCATACAAATAATGGTAATTTATGCAATTTGTCATACCAATTTGTTTCTAATTTAAATGTTTACAAATGAAGAAATTAAGACCTAATAATGATGAAATGATGTTAAATGTAATGCAAAAAATTTGCTAAATCTACACTAGGTCTGAAATGCTGAAATAGAT
GATAATGGTGGATAGACAGTTCAACCTGTACAGCACCTGATTTCTACTAAAGTTGATGTAATCAACATACCTAACCCCATAGGTAACTCTCAAAAGCAATGTTATTAATCTTGGAATGCTGTATCTGAGCTCCAAATTTAATGGTAATAAATAGTTACCAAAACCTC
CAATTAAGCAGGCATTACTAAGAAGAAAAATCATTAATACAGCATGACCAACTACTAAAAATTAATAATGTGAATTACCATGTAAATATTGTGAACCAGGTGCAGCTAATCTAATCTAATGATTAAAGACATTGCTGTTCTGCCATACCCTAAAAATAGCTAACTAAT
AAAAATAATACTGCAATATCTTTTGCAATTTGTTGAATATAATCATCTTTGTACCAT
>Q0060 1248
TTATTTATTTTCTAAATATGTAATTTTTTAAATAGATCAAGTAAAGTAGTTCTAGTTTATTATATAATCTTTAGTGAATTTTATACCTAATTTCTCAACTACCATTAAGTTTTATACCTTTGACCCCTCATTATTAAATTAATAAATACTAATTTAGATCAATGTAAAAATCTAAG
TGTTTAGATGATAATAAGAAATATTTAATAAGTATTTCTAATTACCTTTAATAATTTTATATAGATTAGCCACTTACCTTTATAACCTATAAATATAGTTTATAGGTATTTAAGAGATCTAATAAATTTAAATTAGCTTCTACTATATAAATTAACATTAAATATAGTG
CAATTTGCAGACATAAATAAAAAATAAGAAAAATTAATATTATTATTATAGAAATTTTTTGATAATTTTGTCTTAATCTTAACAATAATAAGGCATTGCTCTACTAGAACGATTTTACCATTATTAAATTAATAGAAAAATTAACCATCTGCATCTGCATACCAGCTAA
TCAAGCGTTTGAACCAATATCTGATGATCTCAATGGTTTAAATTTAATAATTTCTATATTTTTTAAATTTATATGTAGAATTGTGTTGAATTAATAAATTTTATAAATTCAGCACCTCTAACAAATGCTTCAATTTAGGTGTTCTCATATATCCATTAATAATTTT
AATAATGATATACACCTTTTAAATCATGAATAAGTCAATAATACATAATTACGATTAATTTTTTATACACCTTTCCACACTTAGTTAAATTAACATAAATAATTAGCTAATCTCAAACTCTCTAATTTAAATACACCAATTAACGGTCTATATTAGATTTTTTTATTG
AAGATGAATTTTTGAACTAGAAATAGTTCCATCACCTTCAATTAATCCAGCTAAATAAGGTCCCACTTATCATAAATTTAATTTAAATATATTTCTTTATTTCTATTACTTGATTATTATTATTACTTTTCATATCTTTTTGGTTACCAAAACCTCCAATTAAGCAGG
CATTTACTAAGAAGAAAAATCATTAATACAGCATGACCAACTACTAAAAATTAATAATTTGTGAATTACCATGTAAATATTGTGAACCAGGTGCAGCTAATCTAATCTAATGATTAAAGACATTGCTGTTCTGCCATACCCTAAAAATAGCTAACATAAAATATAATATCTT
GCAATATCTTTTGCAATTTGTTGAATATAATCATCTTTGTACCAT
>Q0182 405
CTATTTAAAGGACTTGAACCTTCATCTGTAAGATAACATCTTAAGAGTATCGTGTCTACCAATTCACCAAAATAGCTATTAAATAGATAGTTATTAAAAAATAATATATATATATTTTAAATATAAAAAATATTATTAATATTATAAAATATAAAATAAAT
ATTATAATATAAATAAATAAATAATATTATAATATTATAATTAATGCCAGACGAACTCCTTCGGGTCGCGCCCCCCCCCGGGGCGGGCCGGACTATAAAAAATATTTTAAATTTATTTTTTAAAAATAAATTAAGAAATTTTATTATATCCTTACTCTTT
TTATATTATAATAAAAACTCCTTCGGGGTTCGAGATCCCGTGGCCGGGCCCGGAACTAT

```

For your portfolio

The markers will look for and assess the following:

1. RC.py to see if it has the variable seq and prints out the reverse complementary sequence.
2. **get_mito_gene.py** to test if it can read the original fasta file and create a new file **mito_gene.fa** with mitochondria gene sequences in it.
3. **Mito_RC.py** to test if it can successfully finish the job in the instructions.