Python Programming for Life Scientists

Hanna M. Schilbert & Katharina Sielemann
Genetics & Genomics of Plants, CeBiTec, Bielefeld University

Reminder: ‘Studienleistungen’ – study requirements

- **active participation**
- **Documentation of course exercises**
- **Python project**
  - Write a script
  - Give a presentation about it (10-15min)
  - Deadline for the submission of the python script is **15.04.2021 (1pm)**
  - Presentations will be given on the **16.04.2021 (8am)**

Research

Gene prediction

Functional annotation

De novo genome assembly

De novo transcriptome assembly

Non-canonical splice sites

Genome Research

RNA-Seq analysis

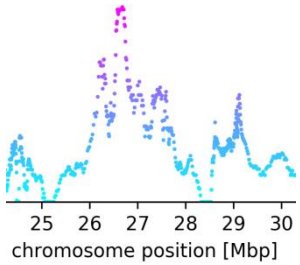
synteny

Co-expression analysis

phylogeny

Gene families / pathways

Mapping-by-sequencing (MBS)



Public scripts: https://github.com/bpucker

Example documentation

```
<html>

    <head>
        <h1>Applied Python Programming for Life Scientists</h1>
        <!-- fill in the following fields -->
        <h2>SEMESTER</h2>
        <h2>NAME</h2>
        <h2>E-MAIL</h2>
    </head>


    <body>


        <!-- here starts one block -->
        <div id="day1">
            <h3>Day 1 - Introduction</h3>
            <p>all important inforamtion</p>
            <p>homework: request UNIX-Account at CeBiTec Support</p>
        </div>
        <!-- here ends one block (can be copied) -->



    </body>

</html>
```

Aims & content

- Learning the basics of Python syntax

```
def print_hello_world():
  """function prints 'hello world!'"""
  print("hello world!")
print_hello_world() #calling function
```

```
hello world!
```

Aims & content

- Learning the basics of Python syntax
- parsing, filtering, exporting, converting

Genbank (.gb)

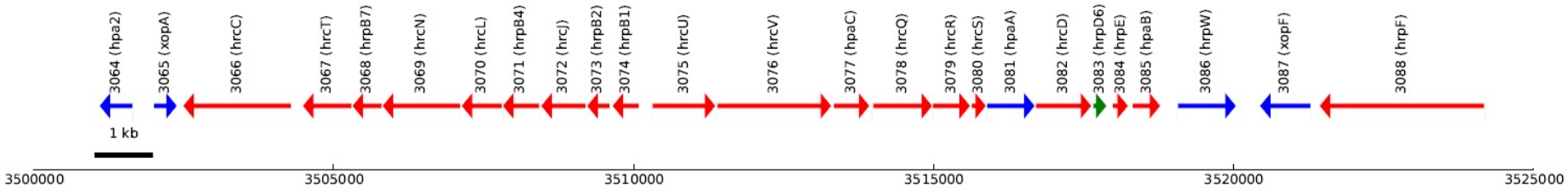
```
          ncRNA                     :: 1
          Frameshifted Genes        :: 35
          ##Genome-Annotation-Data-END##
          COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..5079002
                     /organism="Xanthomonas campestris pv. campestris"
                     /mol_type="genomic DNA"
                     /strain="B100"
                     /db_xref="taxon:340"
                     /pathovar="campestris"
     gene            1..1329
                     /locus_tag="XCCB100_RS00005"
                     /old_locus_tag="xcc-b100_0001"
                     /old_locus_tag="xccb100_0001"
     CDS             1..1329
                     /locus_tag="XCCB100_RS00005"
                     /old_locus_tag="xcc-b100_0001"
                     /old_locus_tag="xccb100_0001"
                     /inference="EXISTENCE: similar to AA
                     sequence:SwissProt:Q8PRG2.1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
                     /product="chromosomal replication initiator protein DnaA"
                     /protein_id="WP_011035259.1"
                     /db_xref="GI:499345720"
                     /translation="MDAWPRCLERLEAEFPPEDVHTWLKPLQAEDRGDSIVLYAPNAF
                     IVEQVRERYLPRIRELLAYFAGNGEVALAVGSRPRAPEPLPAPQAVASAPAAAPIVPF
                     AGNLDSHYTFANFVEGRSNQLGLAAAIQAAQKPGDRAHNPLLLYGSTGLGKTHLMFAA
                     GNALRQANPAAKVMYLRSEQFFSAMIRALQDKAMDQFKRQFQQIDALLIDDIQFFAGK
                     DRTQEEFFHTFNALFDGRQQIILTCDRYPREVEGLEPRLKSRLAWGLSVAIDPPDFET
                     RAAIVLAKARERGAEIPDDVAFLIAKKMRSNVRDLEGALNTLVARANFTGRSITVEFA
                     QETLRDLLRAQQQAIGIPNIQKTVADYYGLQMKDLLSKRRTRSLARPRQVAMALAKEL
                     TEHSLPEIGDAFAGRDHTTVLHACRQIRTLMEADGKLREDWEKLIRKLSE"
     gene            1605..2705
                     /locus_tag="XCCB100_RS00010"
                     /old_locus_tag="xcc-b100_0002"
                     /old_locus_tag="xccb100_0002"
     CDS             1605..2705
                     /locus_tag="XCCB100_RS00010"
                     /old_locus_tag="xcc-b100_0002"
                     /old_locus_tag="xccb100_0002"
                     /EC_number="2.7.7.7"
                     /inference="EXISTENCE: similar to AA
                     sequence:RefSeq:WP_006451791.1"
                     /note="binds the polymerase to DNA and acts as a sliding
                     clamp; Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
```

```
>xccb100_0001   chromosomal replication initiation protein
MDAWPRCLERLEAEFPPEDVHTWLKPLQAEDRGDSIVLYAPNAFIVEQVRERYLPRIREL
LAYFAGNGEVALAVGSRPRAPEPLPAPQAVASAPAAAPIVPFAGNLDSHYTFANFVEGRS
NQLGLAAAIQAAQKPGDRAHNPLLLYGSTGLGKTHLMFAAGNALRQANPAAKVMYLRSEQ
FFSAMIRALQDKAMDQFKRQFQQIDALLIDDIQFFAGKDRTQEEFFHTFNALFDGRQQII
LTCDRYPREVEGLEPRLKSRLAWGLSVAIDPPDFETRAAIVLAKARERGAEIPDDVAFLI
AKKMRSNVRDLEGALNTLVARANFTGRSITVEFAQETLRDLLRAQQQAIGIPNIQKTVAD
YYGLQMKDLLSKRRTRSLARPRQVAMALAKELTEHSLPEIGDAFAGRDHTTVLHACRQIR
TLMEADGKLREDWEKLIRKLSE*
>xccb100_0002   DNA polymerase III subunit beta
MRFTLQREAFLKPLAQVVNVVERRQTLPVLANLLVQVNNGQLSLTGTDLEVEMISRTMVE
DAQDGETTIPARKLFDILRALPDGSRVTVSQTGDKVTVQAGRSRFTLATLPANDFPSVDE
VEATERVAVPEAGLKELMERTAFAMAQQDVRYYLNGLLFDLRDGLLRCVATDGHRLALCE
TELEKSGSAKRQIIVPRKGVTELLRLLEAADRDVELELGRSHIRVKRGDVTFTSKLIDGR
FPDYEAVIPIGADREVKVDREALRASLQRAAILSNEKYRGVRVEVSPGQLKISAHNPEQE
EAQEEIEADTKVDDLAIGFNVNYLLDALSALRDEHVVIQLRDANSSALVREASSEKSRHV
VMPLRL*
>xccb100_0003   recombination protein F
MSTADHVCSAPSDAGLQGQADRSMHVARLSIHRLRRFEAVEFHPASTLNLLTGDNGAGKT
SVLEALHVMAYGRSFRGRVRDGLIRQGGQDLEIFVEWRERAGDSTERTRRAGLRHSGQEW
TGRLDGEDVAQLGSLCAALAVVTFEPGSHVLISGGGEPRRRFLDWGLFHVEPDFLALWRR
YARALKQRNALLKQGAQPQMLDAWDHELAESGETLTSRRLQYLERLQERLVPVATAIAPS
LGLSALTFAPGWRRHEVSLADALLLARERDRQNGYTSQGPHRADWAPLFDALPGKDALSR
GQAKLTALACLLAQAEDFAHERGEWPIMALDDLGSELDRHHQARVIQRLASAPAQVLITA
TELPPGLADAGKTLHRFHVEHGQLVPQPLPTDPPRLA*
>xccb100_0004   DNA gyrase subunit B
MTDEQTTPPTPNGTYDSSKITVLRGLEAVRKRPGMYIGDVHDGTGLHHMVFEVVDNSVDE
ALAGHADDIVVKIHVDGSVAVSDNGRGVPVDIHKEEGVSAAEVILTVLHAGGKFDDNSYK
VSGGLHGVGVSVVNALSEHLWLDIWRDGFHYQQEYALGEPQYPLKQLEASTKRGTTLRFK
PAVEIFSDVEFHYDILARRLRELSFLNSGVKIALIDERGERRDDFHYEGGIRSFVEHLA
QLKTPLHPNVISVTGEHNGIVVDVALQWTDAYQETMYCFTNNIPQKDGGTHLAGFRGALT
RVLSNYIEQNGIAKQAKITLTGDDMREGMIAVLSVKVPDPSFSSQTKEKLVSSDVRPAVE
NAFGARLQEFLQENPNEAKAITGKIVDAARAREAARKARDLTRRKGALDIAGLPGKLADC
QEKDPALSELFIVEGDSAGGSAKQGRNRKNQAVLPLRGKILNVERARFDRMLASDQVGTL
ITALGTGIGRDEYNPDKLRYHRIILMTDADVDGSHIRTLLLTFFYRQMPELIERGYIYIG
LPPLYKLKQGKSELYLKDDAALNAYLASSAVEGAALIPASDEPPITGEALEKLLLLFAGA
KEAIARNAHRYDPALLTALIDLPPLDVVQLQAEGDVHPTLDALQAVLNRGTLGTARYHLR
FDPATDSAAASLVSVRKHMGEEFTQVLPMGAFESGELRPLREVALALHGLVREGAQILRG
NKSHPITSFAQAQAWLLEEAKRGRQVQRFKGLGEMNAEQLWETTVNPDTRRLLQVRIEDA
VAADQIFSTLMGDVVEPRRDFIEDNALKVSNLDI*
>xccb100_0005   putative membrane protease
MSAVLPPSPAPVSVPGPPSLRSAVLGFCIDLLIAIGLLLLLSVAGFAVWGFLRSMGEVQA
VRAQGGSPSPAAIMAAIGQPGVMVQLLIALVSTATPAVLLYFWRRRATPAEQATSRAAIR
RRSTWGWIAAVAAGVFMLSNLVSVLASALGIKPVPTNLPLMEEAIKQWPLALVVFAVAIA
PAYEELLFRRVLFGRLLAAGRPWLGVVLSSLTFALVHEVPGISGNGVVAIAQLWLVYGGM
GAAFAWLYWRTGTLWAPILAHGINNATALAALYFFGLG*
>xccb100_0006   putative exported peptidase/protease
MKVRLLIVVAVLALTACATTTSPTGRRQVVGGVTQDQLDKLGAESFAQTKAKEKVSTDGK
QNAYVQCVVNALVAQLPPQWRETRWETALFVDDEANAFALPGGKVGVNTGIFTVAKTQDQ
LAAVLGHEIGHVISRHHEERITRQLGAQTGLGIIGALAGAAYGDGAASAVNQVGGMTAQT
VFLLPGSRTQESEADVVGQRLMAQAGFDPAQAVSLWQNMMAASGNRQPQWLSTHPDPANR
IRELQADVNALQPVVQQARQDGRVPRCG*
```

FASTA (.fa/.fas/.fasta)

Aims & content

- Learning the basics of Python syntax
- parsing, filtering, exporting, converting
- Generate figures (e.g. matplotlib)

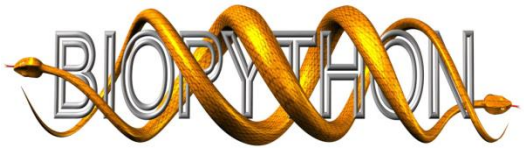


Aims & content

- Learning the basics of Python syntax
- parsing, filtering, exporting, converting
- Generate figures (e.g. matplotlib) 
- Own projects/challenges?! => own solution & success

Why Python?

- Easy to learn
- More efficient than Excel => important for big data
- Check/modify => unlimited opportunities
- Script needs to be written once and can be applied often
  => frequent problems are ideal for bioinformatics
- Very powerful with many libraries/modules
  => biopython (including NCBI BLAST+), scipy, numpy, Rpy, matplotlib…



Examples

- Everything Excel does and much more
- Primer design and validation
  - Identify binding sites and their distances, nucleotide composition, codon frequency, etc.
- BLAST + evaluation of results
  - *in silico* translation of multiple sequences and automatic inspection of all resulting gene products
- Convert data formats (e.g. FASTA-like > FASTA)
- Advanced search&replace
- Your own ideas!!!

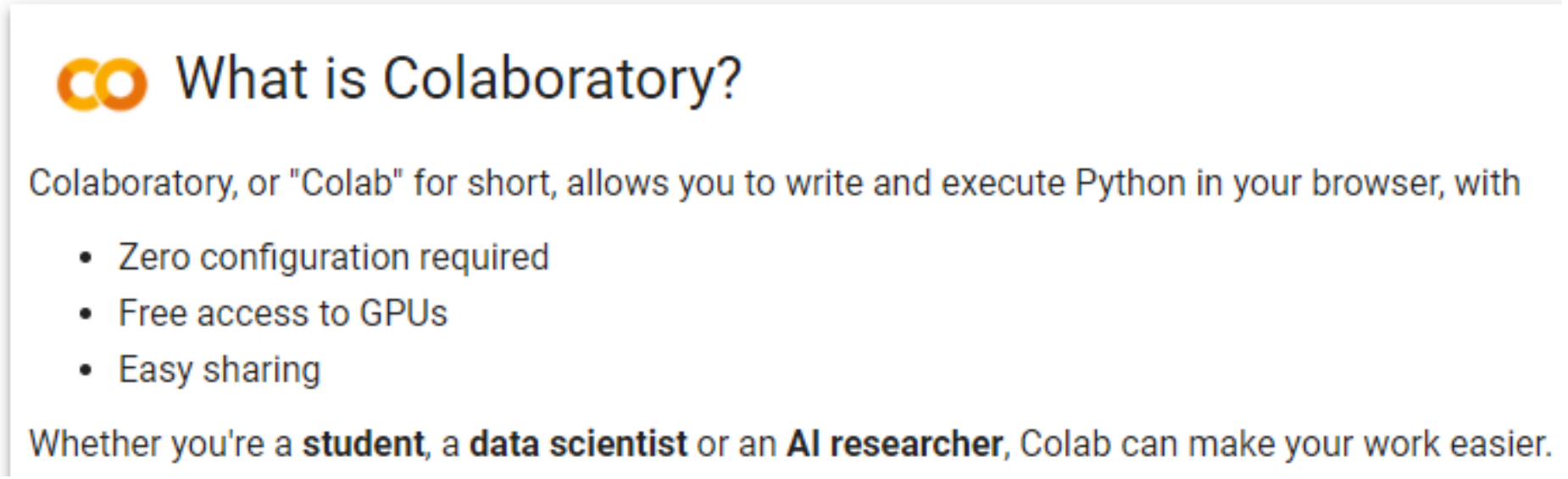
Outline

- Environment: Jupyter notebooks, Python
- Basic commands and data structures
- functions
- Control structures (if, else, for, while)
- File handling
- …
- Your own project!!!

Environment: Jupyter notebooks & Colab

Jupyter notebooks & Colab

- Google account (+ PW) required



Code blocks and text blocks

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ]  seconds_in_a_day = 24 * 60 * 60
     seconds_in_a_day
```

```
86400
```

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter". To edit the code, just click the cell and start editing.

Variables that you define in one cell can later be used in other cells:

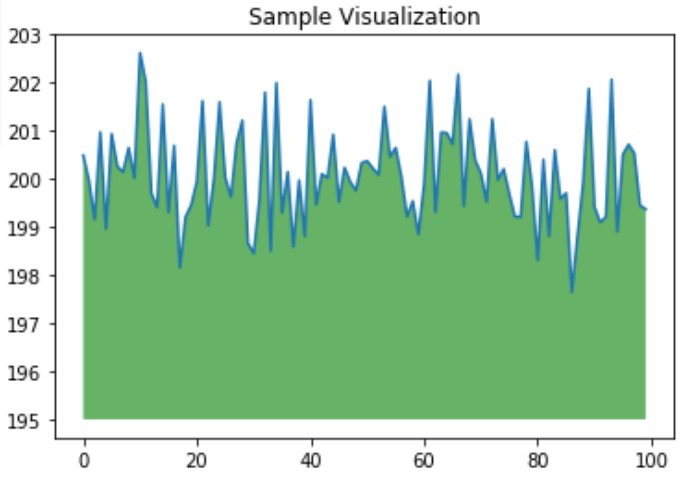
```
[ ]  seconds_in_a_week = 7 * seconds_in_a_day
     seconds_in_a_week
```

```
604800
```

libraries

- combine **executable code** and **rich text** in a single document, along with **images**, **HTML**, **LaTeX** and more
- Use Python libraries (e.g. matplotlib)

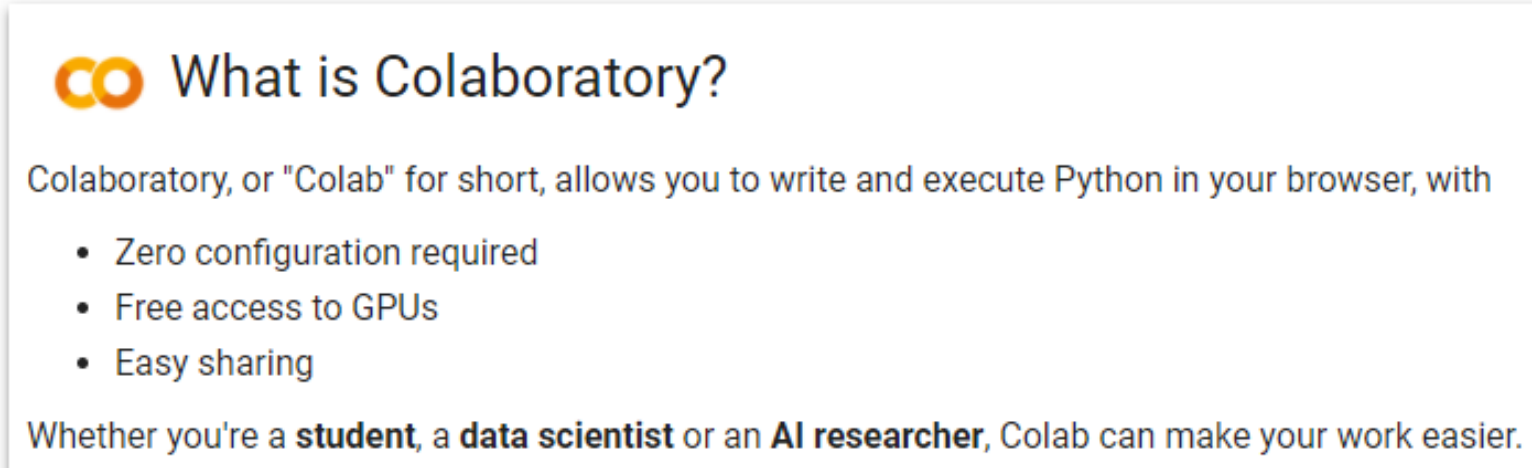
```
[ ]  import numpy as np
     from matplotlib import pyplot as plt

     ys = 200 + np.random.randn(100)
     x = [x for x in range(len(ys))]

     plt.plot(x, ys, '-')
     plt.fill_between(x, ys, 195, where=(ys > 195), facecolor='g', alpha=0.6)

     plt.title("Sample Visualization")
     plt.show()
```



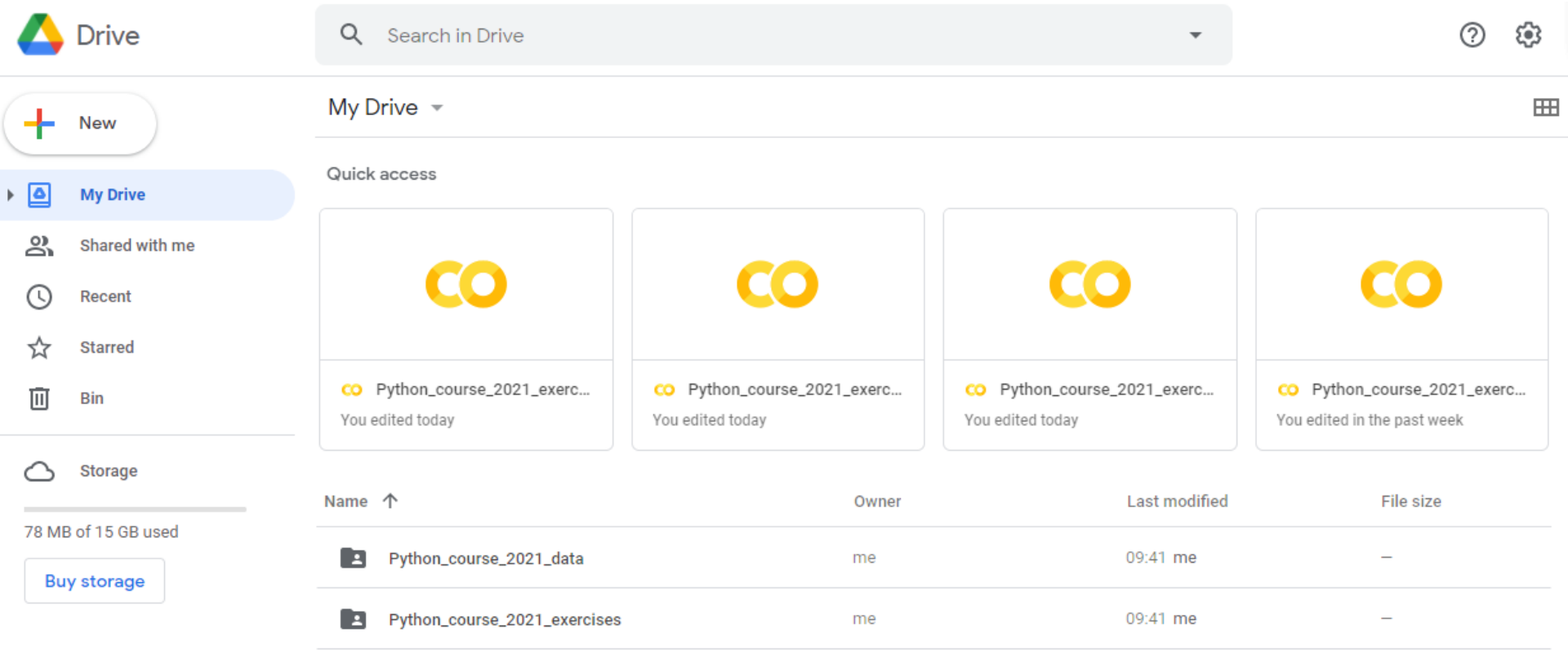
Introduction to Colab

Get started with Google Colaboratory

https://www.youtube.com/watch?v=inN8seMm7UI

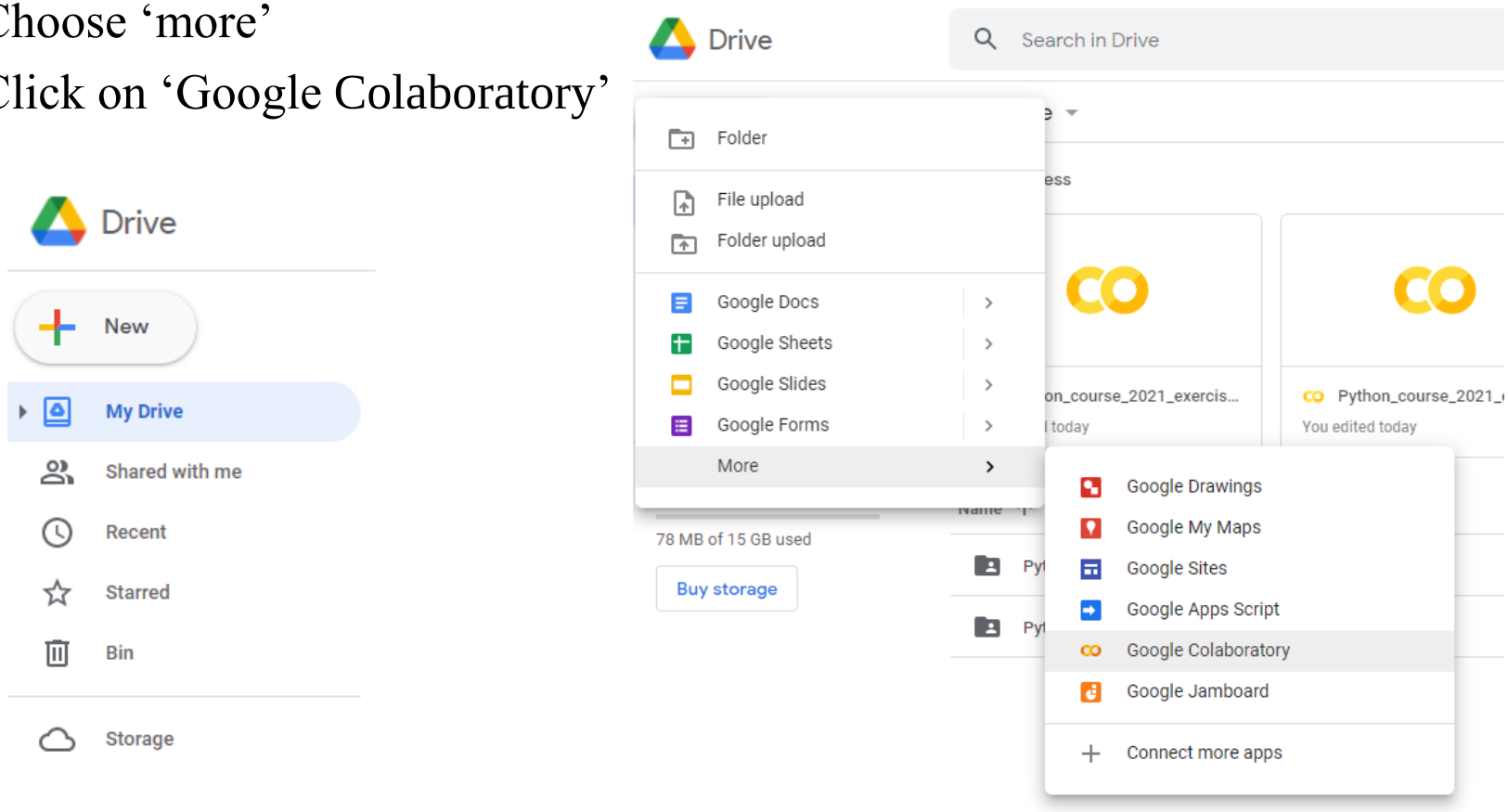


File manager – Google Drive



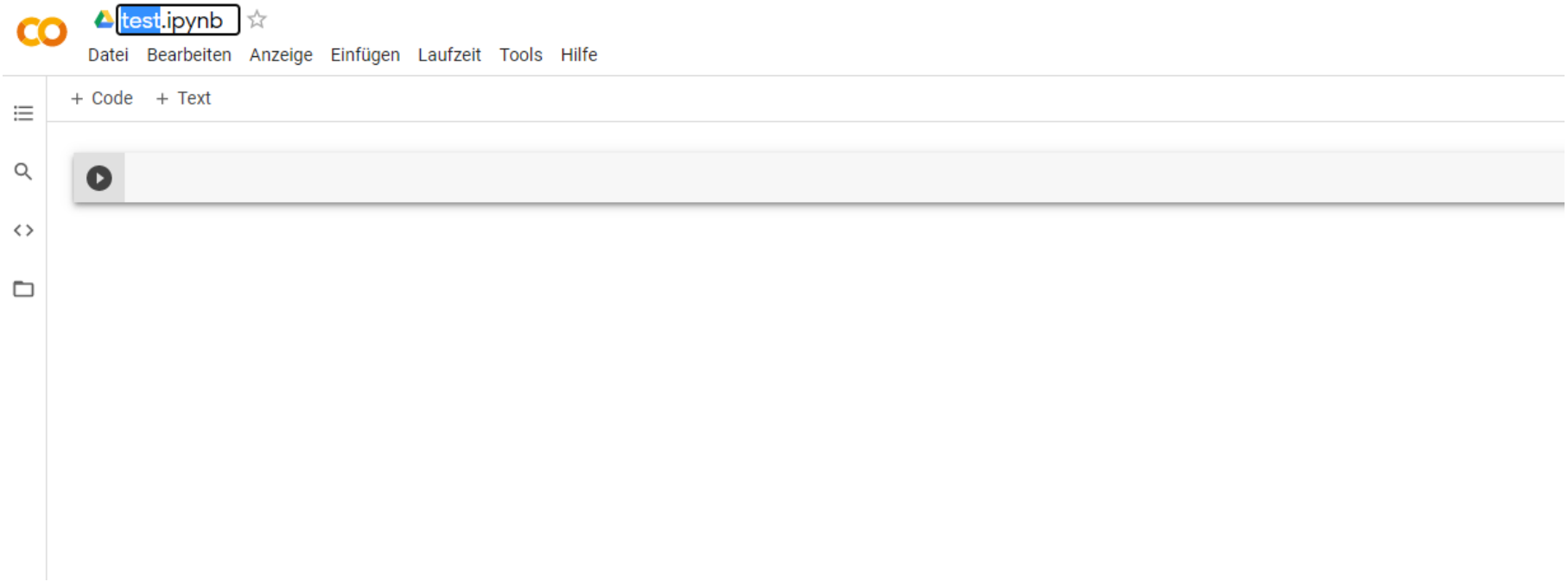
LET'S START - The first Python script

- Create new colab file by clicking on the '+' button
- Choose 'more'
- Click on 'Google Colaboratory'



LET'S START - The first Python script

- Name your file: 'test'

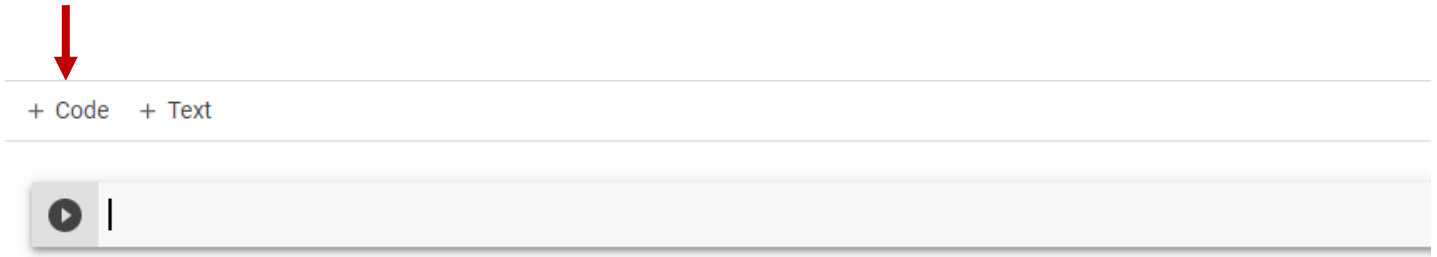


LET’S START - The first Python script

- Delete the existing code block

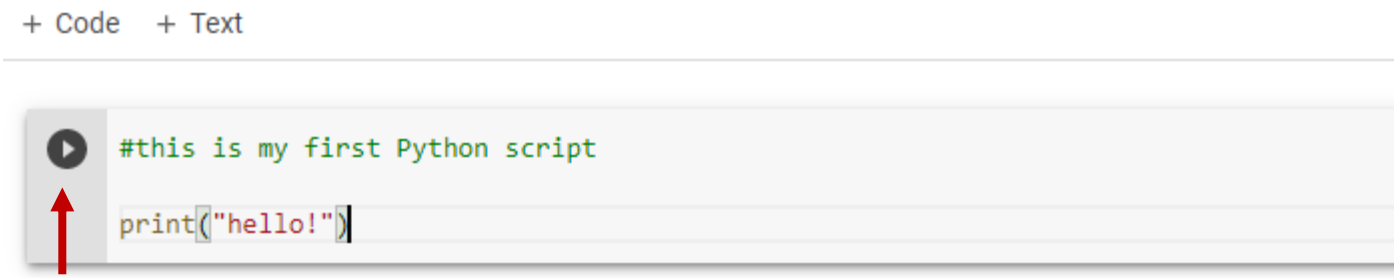
LET'S START - The first Python script

- Add a code block
- Add a comment using #text or "'text"' and write 'this is my first Python script'
- Write 'print("hello!")'



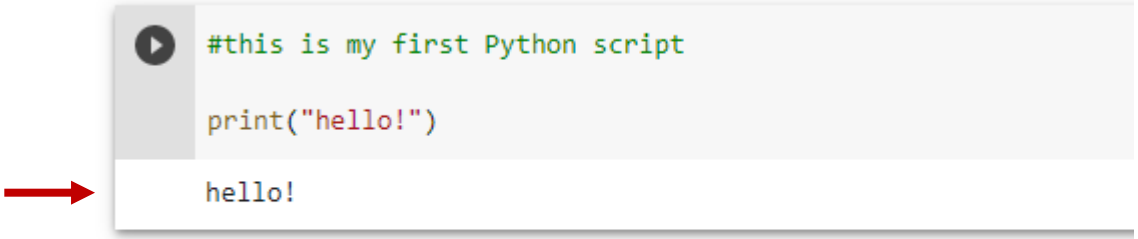
LET’S START - The first Python script

- Run your first script



LET’S START - The first Python script

- Output is directly attached to your code block



Simple commands & variable types

```
1
2  print("test")                                  #print
3  my_list = ["1", "2", "3", 4, 5, 6]  #list
4  my_int = 1                                    #int
5  my_float = float( 3.1 )                       #float
6  my_string = str( my_int )                     #string
7
```

Comments / structure

- Two ways to add comments:
  - '#' rest of the line is comment and ignored by Python
  - ' """ comment """ ' text in triple quotation marks is a comment which can be extended over multiple lines
- Use ASCII characters only (NO ä, ö, ü, ß …)
- Empty lines are ignored by Python

  => Use space to structure code

```
1
2
3   def load_dAFs( filename ):
15
16
17  def load_gene_positions( gff3 ):
30
31
32  def calculate_dAF_per_gene( gene_positions, dAFs ):
51
52
53  def load_variant_effects( filename ):
54      """! @brief load all variant effects from given file """
55
56      variant_effects = {}
57      with open( filename, "r" ) as f:
58          f.readline()  #header
59          line = f.readline()
60          while line:
```

Assignment / comparison

- ‘=’ used to assign value to variable:

  a = “hello world!”

  print(a)

- ‘==’ compares two values/variables:

  a = “hello world!”

  b = “test”

  c = “test”

  a == b

  b == c

- Variable names may contain characters, underline, and numbers (not at the start!)

Variable type string

- a, b, and c are strings (“str”)

- Python allows to check the variable type:

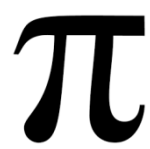
  type(a)

- Almost all variable types can be converted to string:

  str(<VARIABLE>)

Variable types integer & float

- Two variable types for numbers:
  - integer = complete number (example: 3)
  - float = decimal number (example: 3.1415926)

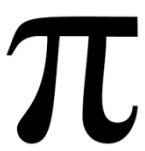


- Important: “ . ” NOT “ , ” separates numbers in float!

- Some strings can be converted to integer/float:

  my_int = int(“3”)

  my_float = float(“3.145926”)



- Check result via type(<VARIABLE>)

Python as calculator

- Numbers can be used for calculations ;-)

```
a = 3
b = 2
print(a+b)          #addition
print(a*b)          #multiplication
print(a**b)         #exponentiation
print(a/b)          #division
print(a/float(b))
print(a%b)          #modulo division
print(a < b)        #alternatives: >=, <=, >, and ==
print(a != b)       #test for inequality
```

- Calculating roots?
- Interested in more complex math? => numpy, scipy

Variable type list

- List can contain elements of different types (e.g. strings):

```
my_list = [ a, b, c ]              #list of strings
print(my_list)
new_list = [ “eins”, “zwei”, “drei” ]
```

- Elements can be accessed via index
- Index is given in square brackets after the list name:

```
print(new_list[ 1 ])
```

- Matching your expectation?

Indices in Python

- Python starts counting at 0!!!

```
new_list = [ “eins”, “zwei”, “drei” ]
#index:          0          1          2
```

- Lists can be concatenated:

```
new_list = new_list + [ “vier”, “fuenf”, “sechs”, “sieben” ]
#new_list = [ “eins”, “zwei”, “drei”, “vier”, “fuenf”, “sechs”, “sieben” ]
#index =        0          1          2          3          4          5              6
```

- Print subset of list:

```
print(new_list[ 3: ])          #elements with index to end
print(new_list[ :3 ])          #elements in front of the given index
print(new_list[ 3:5 ])         #elements with first index in front of second
```

Indices in Python II

- Strings have indices as well:

```
a = “hello world test string!”
print(a[1:])
print(a[5:10])
print(a[:-1])        #-1 = only last element
print(a[-5:-1])      #-5 = starting from 5th element from the end of the
                     #list without the very last one
```

Variable type boolean (True/False)

- Already used for comparison:

```
print(1 == 1)
print(1 > 1)
print(1 == True)
print(True+True)
print(True + False)
```

- Boolean variables can be used for calculations (like numbers)

- Most of the time used only for internal calculations

Brackets

- Two important types of brackets:

  [ ] to generate lists and to access elements via index

  ```
  a = [ ]          #empty list
  b = “test”       #a string
  print(b[1])
  ```

  () to transfer arguments to functions

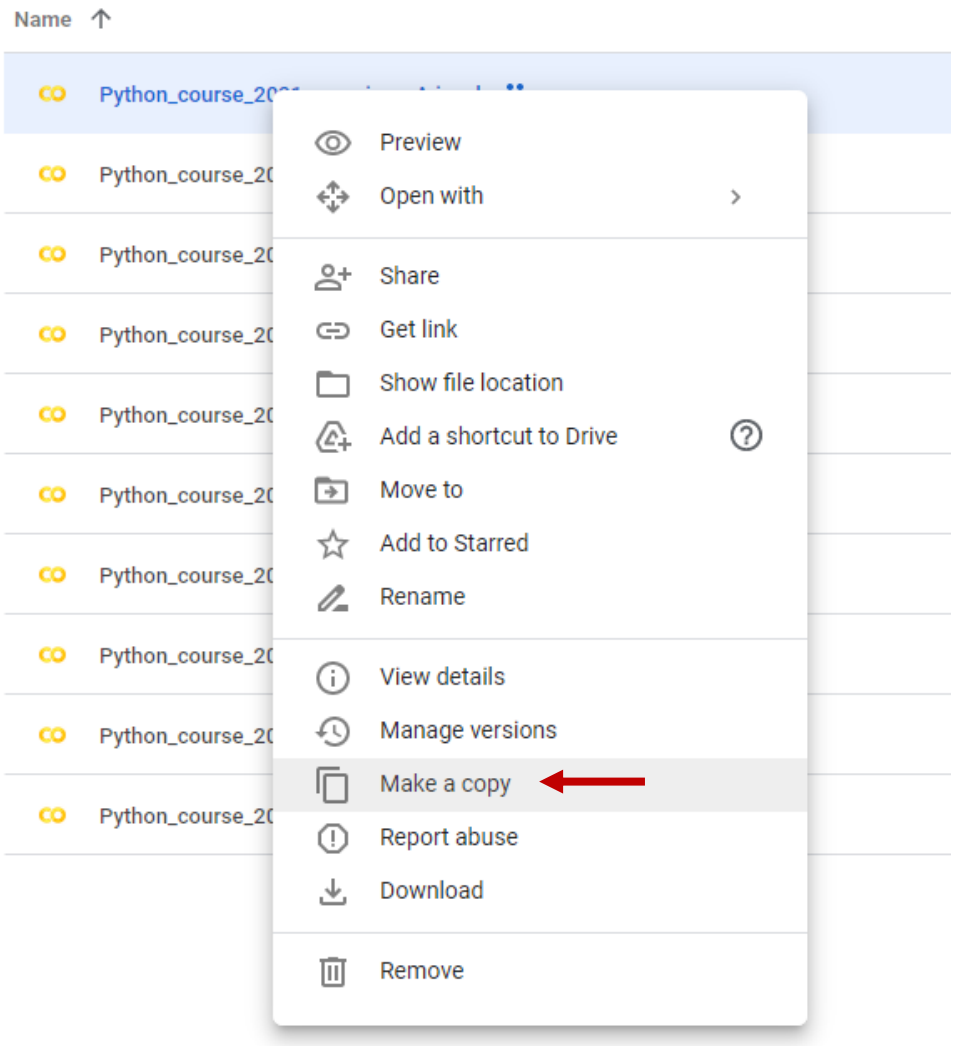
- => What are functions?

  Examples:

  - str(<VARIABLE>)
  - int(<VARIABLE>)
  - float(<VARIABLE>)

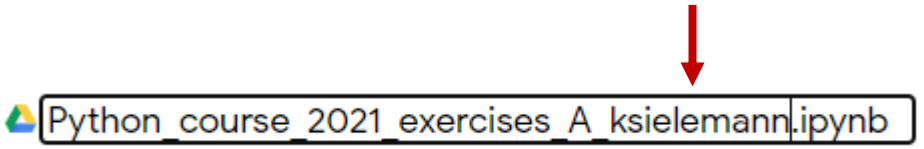
Exercises

- Use the pre-prepared exercise sheets!
- Use the **link** in the chat
- Google Drive should open
- Very **important**!: ‚right click‘ on the file and make a copy!
- Google Colab should open



Exercises

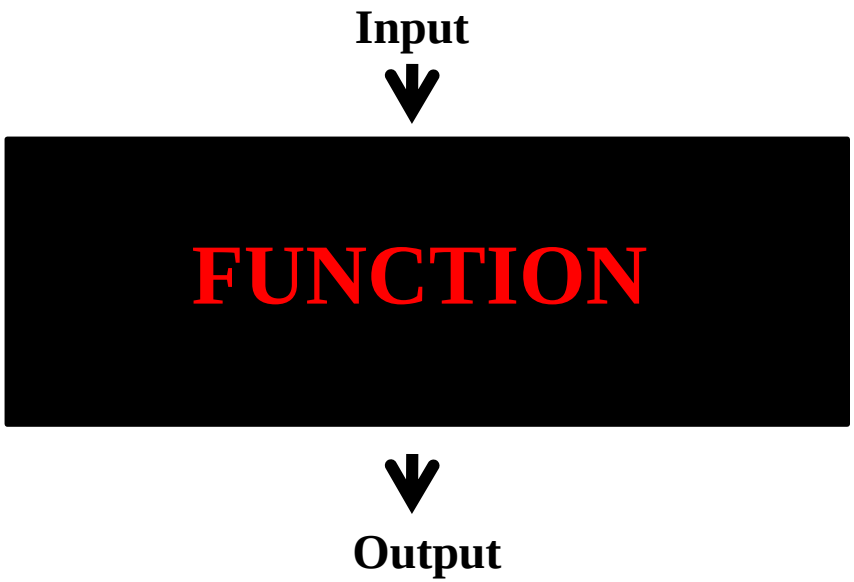
- Rename the file
- Save your own file



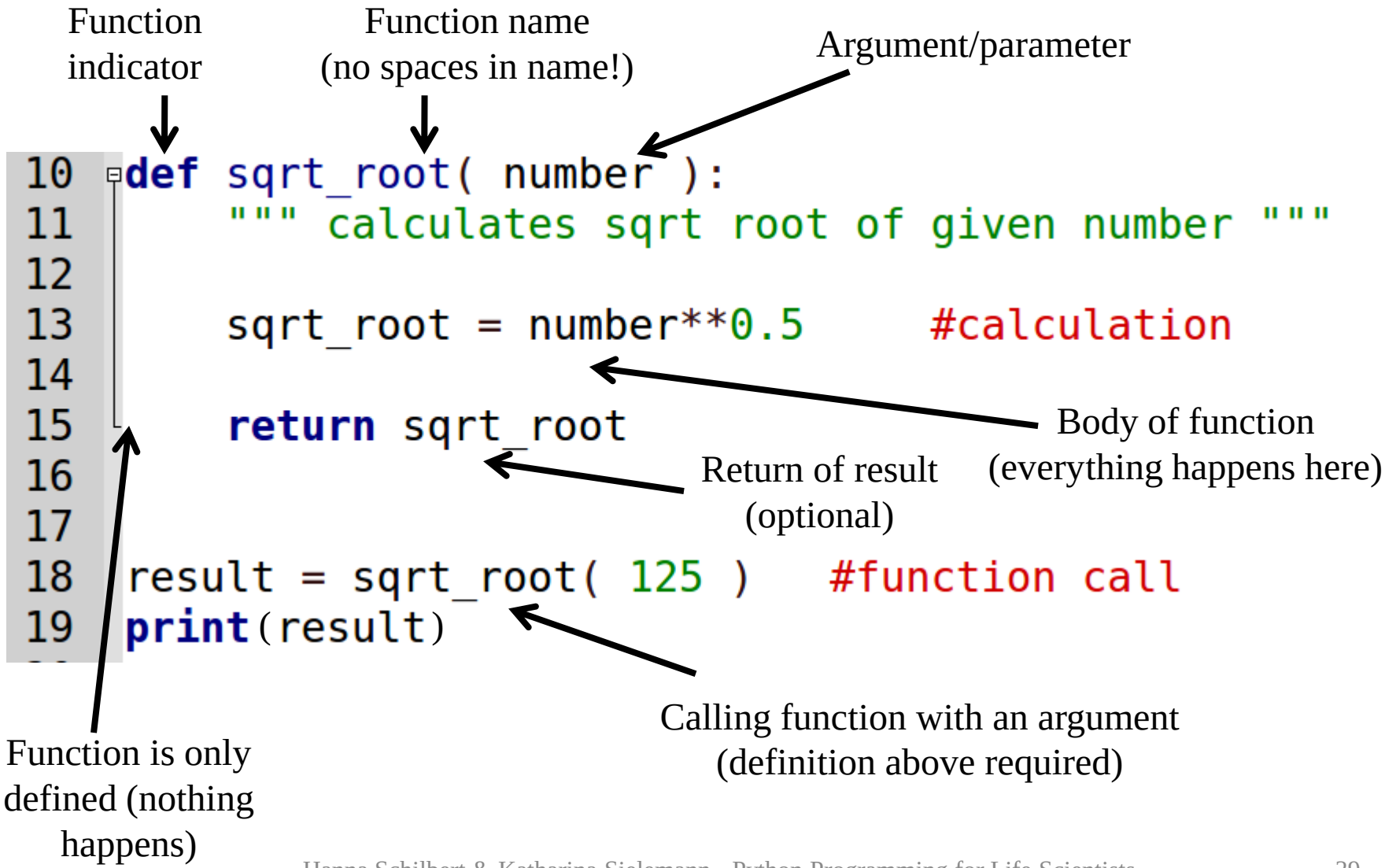
Exercises A – Part1

- 1.1) Save 3,14159265359 in a variable of type float!
- 1.2) Convert variable from float to integer!
- 1.3) Convert variable back! What happens?
- 1.4) Convert variable type to string!
- 1.5) Save ‘Python’ in a string variable!
- 1.6) Convert variable type to float! What happens?
- 1.7) What is a pitfall in regards to division when working with int/float?

Functions



Functions


Function indicator
Function name (no spaces in name!)
Argument/parameter
10 def sqrt_root(number):
11 """ calculates sqrt root of given number """
12
13 sqrt_root = number**0.5 #calculation
14
15 return sqrt_root
16
17
18 result = sqrt_root(125) #function call
19 print(result)
Body of function (everything happens here)
Return of result (optional)
Function is only defined (nothing happens)
Calling function with an argument (definition above required)

Advantages of functions

- Generate modules=> write it ones and apply it often (for different purposes)
- structure=> increases readability of your code
- Nesting of calculations:

```
1  def get_gc_content( seq ):
2      """ calculates GC content of given sequence """
3      return float( seq.count('G')+ seq.count('C') ) / len(seq)
4
5  gc_content = get_gc_content("ATGCGACTCAATGCA")
6  print(gc_content)
```

Important functions

```
str( <VARIABLE> )                          #converts variable to string

int( <VARIABLE> )                          #converts variable to integer

float( <VARIABLE> )                        #converts variable to float

<STRING1>.count( "<STRING2>" )             #counts occurrences of string2 in string1

<LISTE>.count(<LISTENELEMENT>)             #counts occurrences of element in list

len(<STRING/LISTE>)                        #calculates length of string/list
```

- Warning: Functions return error if invalid arguments (e.g. wrong variable type) are given!

Exercises A – Part2

- Primer: “ATGCCATGCATTCGACTACG”
- 2.1) Calculate length of primer and print it!
- 2.2) Get number of Gs and print it!
- 2.3) Write a function to analyze the nucleotide composition of a primer and print it!
- 2.4) Is it a suitable primer? Why (not)?

Control structures

if & else

- Distinguish between two cases:

```
1  a = 5 #define variable
2  #user inputs number:
3  b = input("please enter number!")
4  if b < a:   #if b is smaller than a
5      print("b is smaller than a")
6  else:    #in all other cases
7      print("b is NOT smaller than a")
```

- Action depends on result of comparison

elif

- Distinguish between multiple cases:

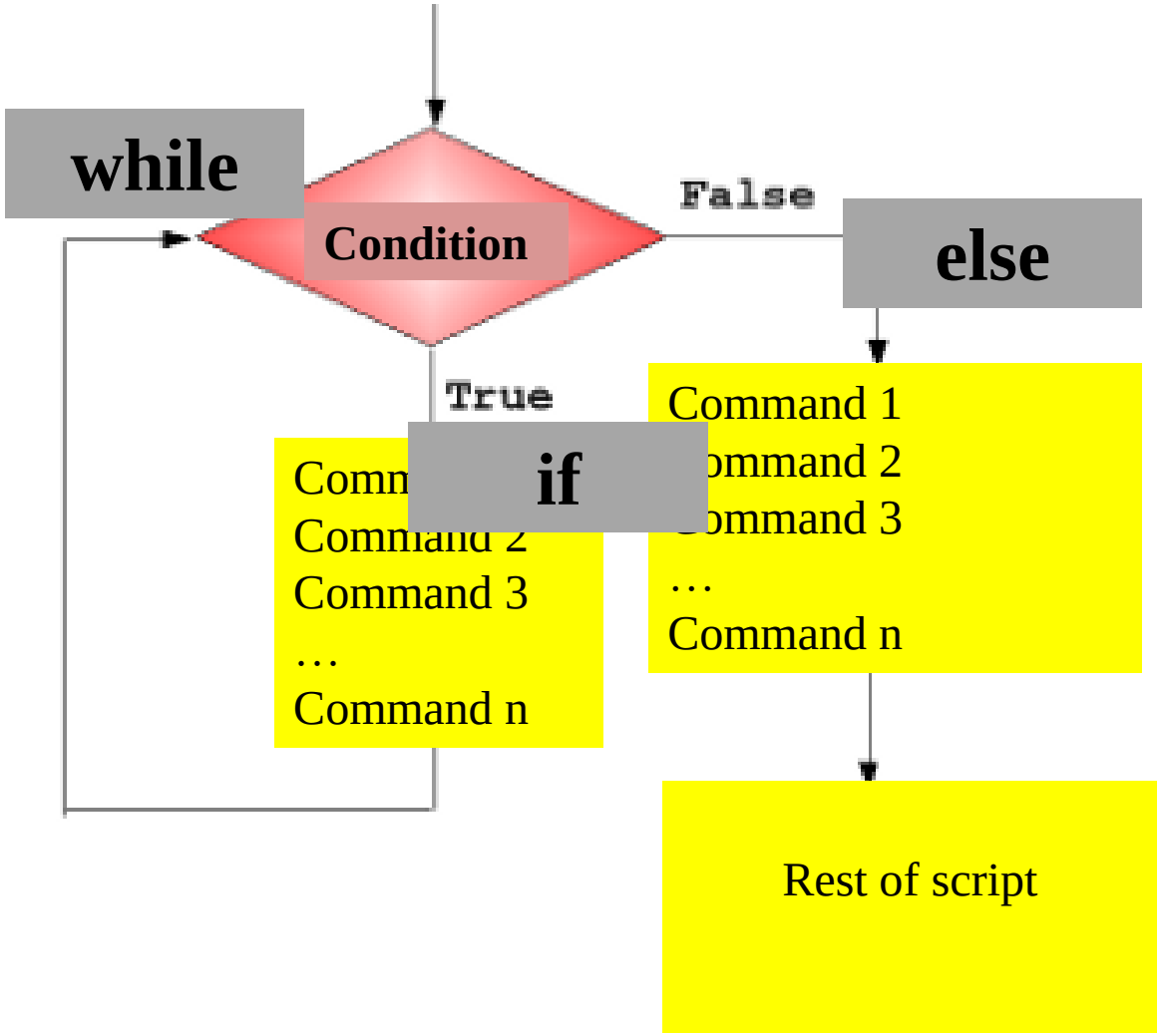
```
1  a = 5 #define variable
2  #user inputs number:
3  b = input("please enter number!")
4  if b < a:   #if b is smaller than a
5      print("b is smaller than a")
6  elif b== a:  #both are equal
7      print("b is matching a")
8  else:    #in all other cases
9      print("b is NOT smaller than a")
```

- Action depends on result of comparison

Exercises B – Part1

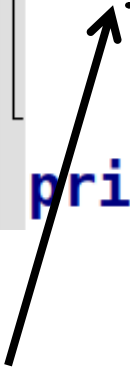
- 1.1) Write script for guessing numbers!
- 1.2) Add tips (smaller/larger) during the guessing process!

Concept of loops



While – loop (example)

```
1  a = 0
2  while a < 10:    #checks if a is smaller than 10
3      print(str( a ) + " is smaller than 10")
4      a += 1  #a = a+1
5      #something useful could happen here
6      print("a was increased by 1")
7  print(str( a ) + "is larger than 10")
```



Code is executed until the condition for this loop becomes false

While – infinite loop

```
1  #infinite loop:
2  a = 0
3  while True: #always true
4      a += 1  #a = a+1
5      print(str( a ))
6  print("this line is never reached")
```

WARNING: this loop is infinite!

For loop

```
1  list_of_species = [ "E.coli", "B.subtilis", "S.cerevisiae", "C.glutamicum", "A.tumefaciens" ]
2  for species in list_of_species:
3      if len( species ) < 12:  #Length of names is calculated and compared
4          print(species)       #Name is printed
5
6  #Line 3+4 is executed several times:
7  #1: species = "E.coli"
8  #2: species = "B.subtilis"
9  #3: species = "S.cerevisiae"
10 # ...
```

For loop

Control variable (species)

List of data (list_of_species)

```
1  list_of_species = [ "E.coli", "B.subtilis", "S.cerevisiae", "C.glutamicum", "A.tumefaciens" ]
2  for species in list_of_species:
3      if len( species ) < 12:  #Length of names is calculated and compared
4          print(species)       #Name is printed
5
6  #Line 3+4 is executed several times:
7  #1: species = "E.coli"
8  #2: species = "B.subtilis"
9  #3: species = "S.cerevisiae"
10 # ...
```

Species name is printed if shorter than 12 characters

Exercises B – Part2

- 2.1) Write a function counting to 100 and printing all number which can be divided by 4 without any residue!
  - Info: 10 % 2 #modulo division in Python
- 2.2) Write a function counting down from 1000 to 0 and printing all numbers!
- 2.3) Generate a list of species names! Write a function printing all species names starting with “E”!
- 2.4) Expand this function to limit the printing to species names which are additionally shorter than 10 characters!
- 2.5) Expand this function to limit the printing to species names which are additionally ending with “a”.

range()

```
1  list_of_species = ["E.coli", "B.subtilis", "S.cerevisiae", "C.glutamicum", "A.tumefaciens"]
2  length = len( list_of_species ) #length = 5
3  for i in range( length ):   #starts at 0 and runs to i=4 (five values)
4      if len( list_of_species[ i ] ) < 12:    #length of name is calculated and compared
5          print(list_of_species[ i ]) #name is printed
6
7  #i is taking five different values:
8  #1: i=0
9  #2: i=1
10 #3: i=2
11 #4: i=3
12 #5: i=4
13 #i=5 is never reached by range()
```

enumerate()

```
1  list_of_species = ["E.coli", "B.subtilis", "S.cerevisiae", "C.glutamicum", "A.tumefaciens"]
2  for idx, species in enumerate( list_of_species ):
3      if len( species ) < 12: #length of names is calculated and compared
4          print("position of " + species + " is: " + str( idx ))
5
6  #i is taking five different values:
7  #1: i=0 and species="E.coli"
8  #2: i=1 and species="B.subtilis"
9  #3: i=2 and species="S.cerevisiae"
10 #4: i=3 and species="C.glutamicum"
11 #5: i=4 and species="A.tumefaciens"
12 #i=5 is never reached
```

Exercises B – Part3

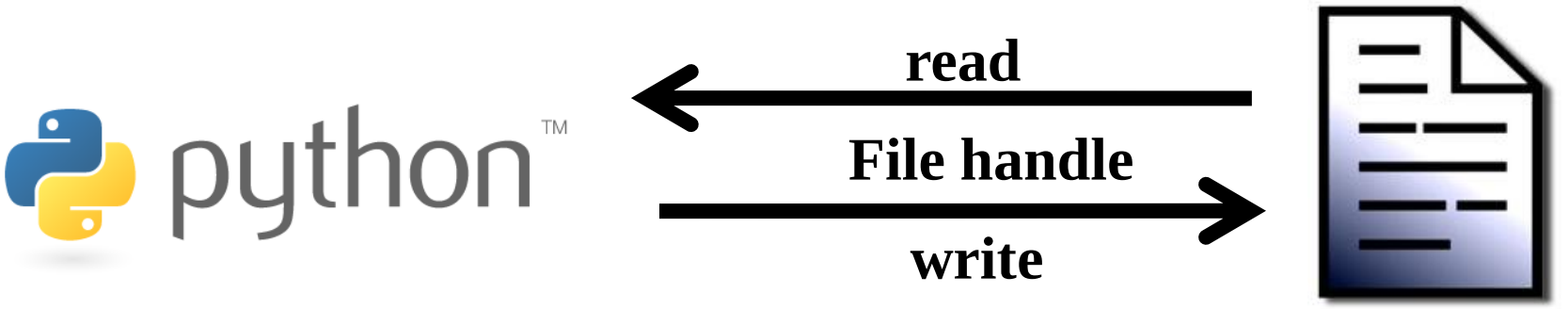
- 3.1) Write a script to print 50x “here” and the current value of the control variable!

- 3.2) Write a script to walk through the species list and to print the character from the species where the index corresponds to the current control variable value!

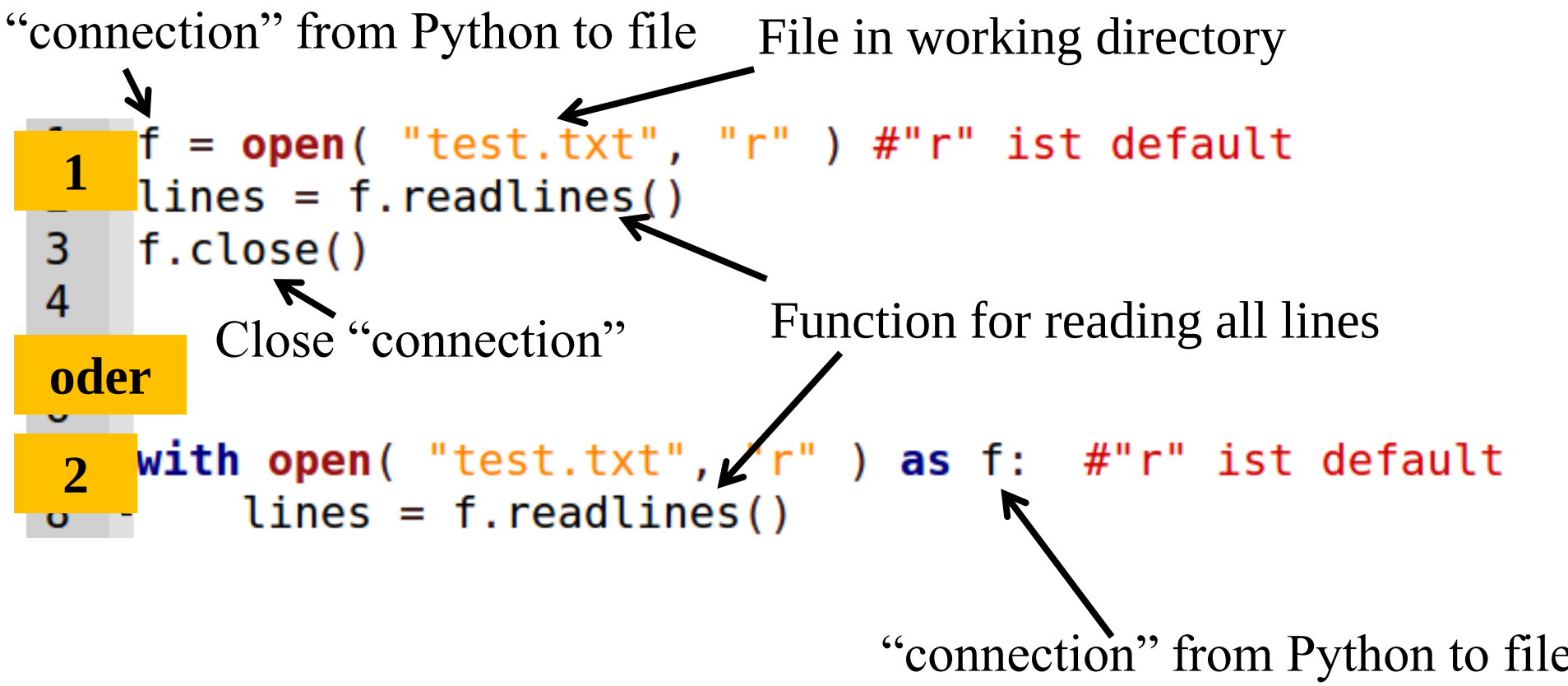
File handling

Concept of file handling

- “connection” from Python to file
- Read = Transfer of data **from** file
- Write = Transfer of data **into** file



Read a file (parsing)



Reading a file (big data)

```
1 with open( "text.txt", "r" ) as f:  #"r" is default
2     line = f.readline() #reads next line
3     while line:   #until end of file is reached
4         print(line)
5         line = f.readline() #reads next line
```

- Advantage: only one line is read and processed at a time
- NGS data (e.g. FASTQ/SAM/BAM/VCF) are usually several GB in size => RAM limitations
- Very long sequence (e.g. genome sequences) in FASTA might be to large for available RAM

Analyze file - example

- How many lines are in AtCol0_Exons.fasta? (large file!)
- Under UNIX: head <DATEINAME>

```
>AT1G01010.1|exon-1 | 1-283 | chr1:3631-3913 FORWARD LENGTH=283
AAATTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAGAGCGAGAGAGATCGAC
GGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAAAATGGAGGATCAAGTTGGGTTTGGGTTCCGTC
CGAACGACGAGGAGCTCGTTGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTGAAGTAGCCATC
AGCGAGGTCAACATCTGTAGCTACGATCCTTGGAACTTGCGCT
>AT1G01010.1|exon-2 | 366-646 | chr1:3996-4276 FORWARD LENGTH=281
TCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTCTCGTAGAGAAAACAACAAAGGGAATCGACAGAGC
AGGACAACGGTTTCTGGTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGG
CTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGTGTTCCTCGATGGAAGATACCCTGACAAAACCAAATCTGATTGGG
TTATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAG
>AT1G01010.1|exon-3 | 856-975 | chr1:4486-4605 FORWARD LENGTH=120
AGGACATATGTCATCTGCAGACTTGAGTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAGATCCCACTCC
CGCTTTTGTCCCCAATATGACTAGTAGTGCAGGTTCTGTG
```

(multiple) FASTA

Name of sequence (header): line starts with '>'

```
>AT1G01010.1|exon-1 | 1-283 | chr1:3631-3913 FORWARD LENGTH=283
AAATTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAGAGCGAGAGAGATCGAC
GGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAAAATGGAGGATCAAGTTGGGTTTGGGTTCCGTC
CGAACGACGAGGAGCTCGTTGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTGAAGTAGCCATC
AGCGAGGTCAACATCTGTAGCTACGATCCTTGGAACTTGCGCT
>AT1G01010.1|exon-2 | 366-646 | chr1:3996-4276 FORWARD LENGTH=281
TCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTCTCGTAGAGAAAACAACAAAGGGAATCGACAGAGC
AGGACAACGGTTTCTGGTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGG
CTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGTGTTCCTCGATGGAAGATACCCTGACAAAACCAAATCTGATTGGG
TTATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAG
>AT1G01010.1|exon-3 | 856-975 | chr1:4486-4605 FORWARD LENGTH=120
AGGACATATGTCATCTGCAGACTTGAGTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAGATCCCACTCC
CGCTTTTGTCCCCAATATGACTAGTAGTGCAGGTTCTGTG
```

(multiple) FASTA

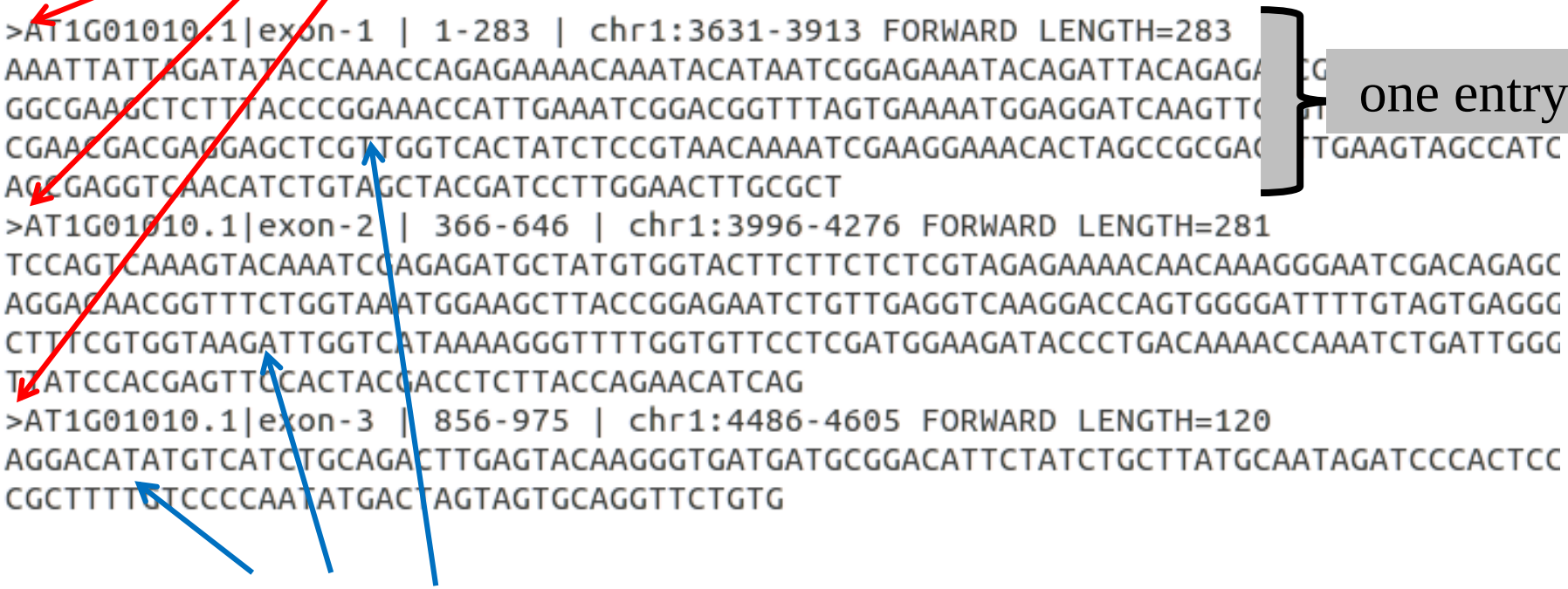
Name of sequence (header): line starts with '>'

```
>AT1G01010.1|exon-1 | 1-283 | chr1:3631-3913 FORWARD LENGTH=283
AAATTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAGA
GGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAAAATGGAGGATCAAGTTC
CGAACGACGAGGAGCTCGTTGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGAC
AGCGAGGTCAACATCTGTAGCTACGATCCTTGGAACTTGCGCT
>AT1G01010.1|exon-2 | 366-646 | chr1:3996-4276 FORWARD LENGTH=281
TCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTCTCGTAGAGAAAACAACAAAGGGAATCGACAGAGC
AGGACAACGGTTTCTGGTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGG
CTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGTGTTCCTCGATGGAAGATACCCTGACAAAACCAAATCTGATTGGG
TTATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAG
>AT1G01010.1|exon-3 | 856-975 | chr1:4486-4605 FORWARD LENGTH=120
AGGACATATGTCATCTGCAGACTTGAGTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAGATCCCACTCC
CGCTTTTGTCCCCAATATGACTAGTAGTGCAGGTTCTGTG
```

one entry

(multiple) FASTA

Name of sequence (header): line starts with '>'



Analyze file - example

```
1 with open( "/vol/apbiokurs/data/AtCol0_Exons.fasta", "r" ) as f:
2     line = f.readline() #reading first line
3     line_counter = 0   #counting lines
4     while line:
5         line_counter += 1    #counting lines
6         line = f.readline()
7 #number in line_counter needs to be converted to string:
8 print("File contains " + str( line_counter ) + " lines")
```

Exercises C – Part1

- 1.1) Count number of sequences (= number of headers) in /vol/apbiokurs/data/AtCol0_Exons.fasta! (add file to your Drive: **link** in chat!)
- 1.2) Count number of sequence lines!
- 1.3) Count number of characters in document! (How many per line?)
- 1.4) How long are all contained sequences combined?
- 1.5) Calculate the average sequence length in this file!

And back again… writing into file!

Read:

```
1
2  with open( "test.txt", "r" ) as f:  #"r" (read) ist default
3      lines = f.readlines()
4
5
6
7
8  Write:
9  with open( "test2.txt", "w" ) as out:
10     out.write( "hello world!" )
```

difference: r = read; w = write

Writes a string into a file

- If output file does not exists, it will be created!
- File handle (f and out) can have any name!

Read & write

```
1 with open( "test.txt", "r" ) as f:  #open file to read
2     with open( "test2.txt", "w" ) as out:   #open file to write
3         line = f.readline() #read from first file
4         while line:
5             #this would be the place to apply filters
6             out.write( line )
7             line = f.readline()
```

Exercises D – Part1

- 1.1) Read the file AtCol0_Exons.fasta and write all headers (starting with '>') into a new file!

- 1.2) Read the file AtCol0_Exons.fasta and write the following:
  - Line if it is a header
  - Length of line if it is a sequence line

- 1.3) Calculate the number of sequences, the cumulative length, and the average length in the new file! Are they matching the values of the original file?

- 1.4) Write sequences into a new file if their length is a multiple of 10!

White space characters

- New line (‘\n’) und tab (‘\t’) are special characters

  print “hello\tworld!\nhello\tworld!\n”

- Python interprets these characters in print statements, but functions like readline() and write() do not!

  => New line needs to be added “manually” to each new line

```
1 with open( "test_file.txt", "w" ) as out:
2     out.write( "first test" )
3     out.write( "second test" )
4     out.write( "third test\n" )
5     out.write( "fourth test\n" )
6     out.write( "fifth test" )
7     out.write( "sixth test" )
```

How many lines does this file have?

strip()

- Removes white space characters from borders of a string (often used for new lines at the line end):

```
1  line = ">name_of_first_seq\n"
2  print( line )
3  #>name_of_first_seq
4  # [empy line genereated by \n ]
5  line = line.strip()
6  print( line )
7  #>name_of_first_seq
```

split()

- Separates a string at each given occurrence of the given substring (e.g. tab, comma, …)

- Generates list of strings

```
1  #tab-delimited file
2  line = "spalte1\tspalte2\tspalte3\tspalte4\n"
3  #line should be splitted at tabs
4  columns = line.strip().split('\t')
5  print (columns)
6  #["spalte1", "spalte2", "spalte3", "spalte4"]
```

join()

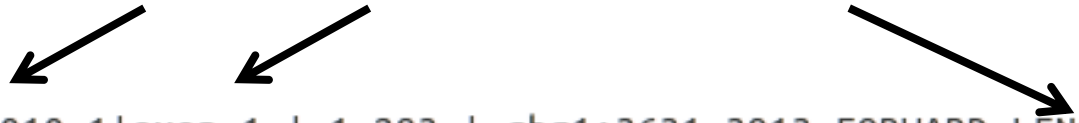
- Combines strings of a list by putting a given substring between them (e.g. underline)

- Important: all elements of list need to be strings!

```
1  #tab-delimited file
2  line = "spalte1\tspalte2\tspalte3\tspalte4\n"
3  #line should be splitted at tabs
4  columns = line.strip().split('\t')
5  print(columns)
6  #["spalte1", "spalte2", "spalte3", "spalte4"]
7
8  new_line = "_".join( columns )
9  print(new_line)
10 #spalte1_spalte2_spalte3_spalte4
```

Exercises D – Part2

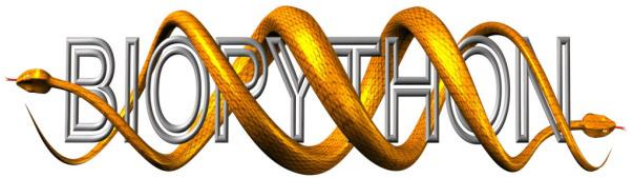
- 2.1) Read the file AtCol0_Exons.fasta and write the following:
  - Only ArabidopsisGeneIdentifier (e.g. AT1G01010)
  - Gene identifier, exon name, and exon length (tab-delimited)



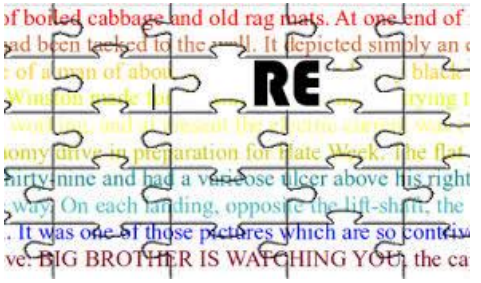
```
>AT1G01010.1|exon-1 | 1-283 | chr1:3631-3913 FORWARD LENGTH=283
AAATTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAGAGCGAGAGAGATCGAC
GGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAAAATGGAGGATCAAGTTGGGTTTGGGTTCCGTC
CGAACGACGAGGAGCTCGTTGGTCACTATCTCCGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTGAAGTAGCCATC
AGCGAGGTCAACATCTGTAGCTACGATCCTTGGAACTTGCGCT
>AT1G01010.1|exon-2 | 366-646 | chr1:3996-4276 FORWARD LENGTH=281
TCCAGTCAAAGTACAAATCGAGAGATGCTATGTGGTACTTCTTCTCTCGTAGAGAAAACAACAAAGGGAATCGACAGAGC
AGGACAACGGTTTCTGGTAAATGGAAGCTTACCGGAGAATCTGTTGAGGTCAAGGACCAGTGGGGATTTTGTAGTGAGGG
CTTTCGTGGTAAGATTGGTCATAAAAGGGTTTTGGTGTTCCTCGATGGAAGATACCCTGACAAAACCAAATCTGATTGGG
TTATCCACGAGTTCCACTACGACCTCTTACCAGAACATCAG
>AT1G01010.1|exon-3 | 856-975 | chr1:4486-4605 FORWARD LENGTH=120
AGGACATATGTCATCTGCAGACTTGAGTACAAGGGTGATGATGCGGACATTCTATCTGCTTATGCAATAGATCCCACTCC
CGCTTTTGTCCCCAATATGACTAGTAGTGCAGGTTCTGTG
```

Modules

Concept of modules



NcbiblastnCommandline
NCBIXML



re.findall()
re.match()

plt.show()
plt.color()

np.array()

Import of modules

```
1  #basic import:
2  import re
3  #import of module under abbreviation:
4  import numpy as np
5  #import part of module:
6  from datetime import datetime
7
8  #usage of module functions:
9  re.findall()
10 datetime.now()
11
12 #try this (requires import):
13 print(str( datetime.now() ))
```

Run time calculation

- Current time is saved in two different places
- Difference is calculated to get the run time

```
1  from datetime import datetime
2
3  t1 = datetime.now()
4  #something should happen here
5  t2 = datetime.now()
6
7  print("it took " + str( t2-t1 ))
```

Regular expressions

- regular expressions (= re) enable efficient search for substrings in a given string

```
1  import re
2  some_string = "AT2G12340.1|exon-1|23745-23965|AT2G12340.2exon-1_23745-23965"
3  hits = re.findall( "AT\dG\d{5}", some_string )  #generates list of hits
4  #searches for "AT\dG\d{5}"
5  #AT, G are matching the very same character
6  #\d is matching all number 0-9
7  #{5} specifies five repetitions of the previous element
8
9  print(hits)
```

Exercises E – Part1

- 1.1) Write all AGIs of AtCol0_exons.fasta into a new file!
- 1.2) Some IDs occur multiple times. Add a filter step to reduce the results to unique IDs!
- 1.3) Calculate frequency of each AGI and construct a histogram (matplotlib)!

  Tip: plt.hist(<LIST_OF_VALUES>)

DNA, RNA, and peptide sequences

Reverse complement

What happens here?

```
1  def revcomp( seq ):
2
3      seq = seq.lower()
4
5      #key:value (=dictionary)
6      complement = { 'a':'t', 't':'a', 'c':'g', 'g':'c' }
7
8      new_seq = []
9
10     for nt in seq:
11         new_seq.append( complement[ nt ] )
12
13     #list[::-1] inverts list (last element becomes first)
14     new_seq = "".join( new_seq[::-1] )
15
16     return new_seq
```

Sequence of bases e.g. ATGACATGA

Converts input to lower case: atgacatga

Get complement for each base

Inverts list (=reverse)

Exercises F – Part1

- 1.1) Write a function to get the reverse complement (upper case letters) of a DNA sequence given in upper case letters!

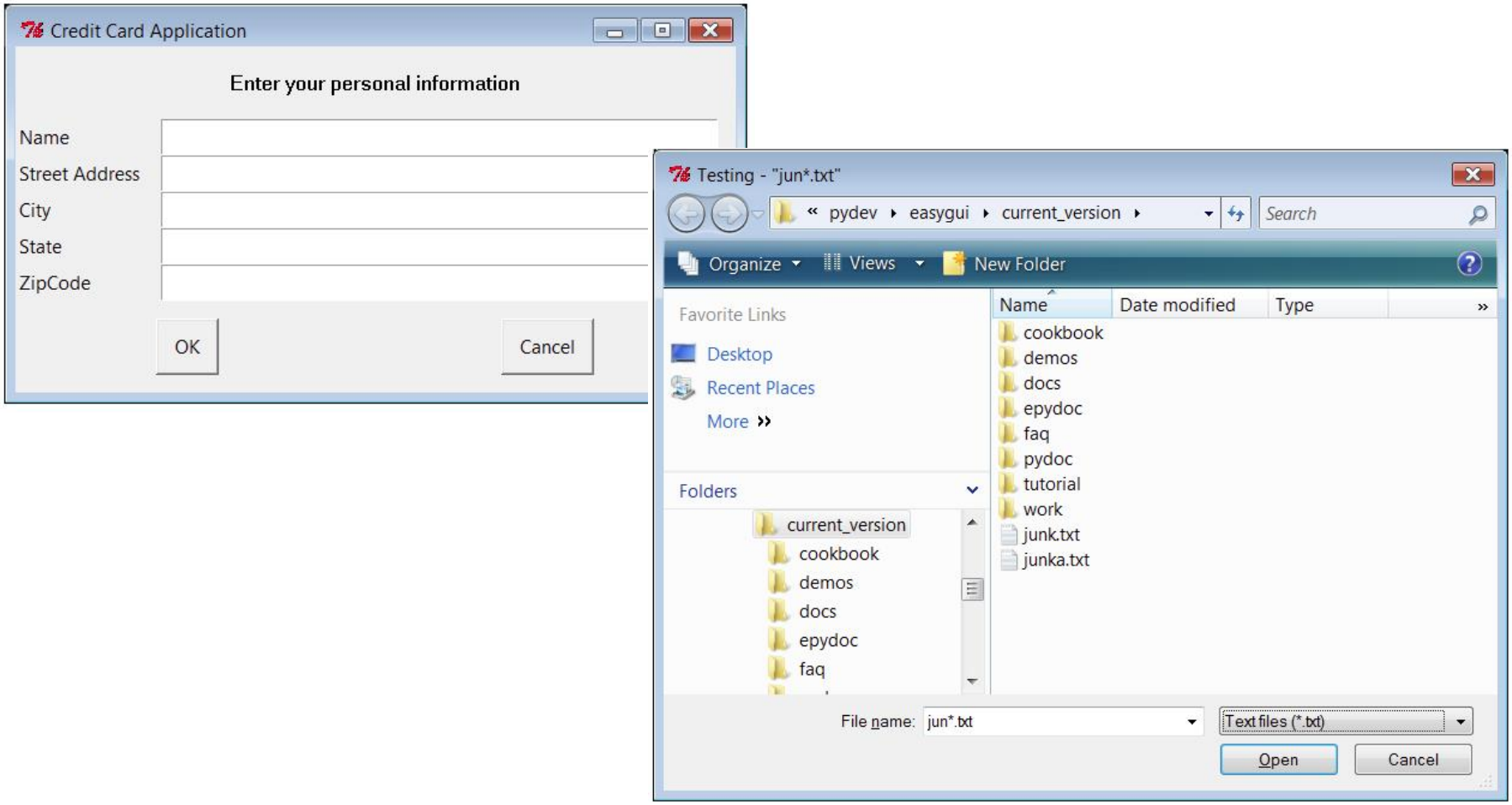
- 1.2) Write a function to convert a DNA sequence into a RNA sequence!

- 1.3) Write a function to translate a DNA sequence into amino acids (first frame only)!

  Tipp: http://en.wikipedia.org/wiki/DNA_codon_table

- 1.4) Write a function to translate DNA sequences in all 6 frames into peptide sequences! The longest peptide sequence per DNA sequence should be returned!

EasyGUI


Credit Card Application
Enter your personal information
Name
Street Address
City
State
ZipCode
OK
Cancel
Testing - "jun*.txt"
« pydev ▸ easygui ▸ current_version ▸
Search
Organize
Views
New Folder
Favorite Links
Desktop
Recent Places
More »
Folders
current_version
cookbook
demos
docs
epydoc
faq
Name
Date modified
Type
cookbook
demos
docs
epydoc
faq
pydoc
tutorial
work
junk.txt
junka.txt
File name: jun*.txt
Text files (*.txt)
Open
Cancel

EasyGUI - documentation

http://www.ferg.org/easygui/tutorial.html

http://easygui.sourceforge.net/tutorial.html#introduction

https://easygui.wordpress.com/

Downloads:
https://pypi.python.org/pypi/easygui/0.97.4#downloads

Two issues

Easygui project shuts down

Posted on 2013/03/06

Effective March 6, 2013, I am shutting down the EasyGui project.

The EasyGui software will continue to be available at its current location, but I will no longer be supporting, maintaining, or enhancing it.

The reasons for this decision are personal, and not very interesting. I'm older now, and retired. I no longer do software development, in any programming language. I have other interests that I find more compelling. I spend time with my family. I play and promote petanque. Life is good, but it is different.

During the course of my software development career I've had occasion to shut down a number of projects. On every occasion when I turned over a project to a new owner, the results were disappointing. Consequently, I have decided to shut down the EasyGui project rather than to try to find a new owner for it.

The EasyGui software will remain frozen in its current state. I invite anyone who has the wish, the will, the energy, and the vision to continue to evolve EasyGui, to do so. Copy it, fork it, and make it the basis for your own new work.

— Steve Ferg, March 6, 2013

Warning about using EasyGui with IDLE

… but it should work!

Selected EasyGUI functions

- msgbox = displays message box
- ccbox / ynbox = poll with yes/no box
- buttonbox = diverse buttons displayed (suitable as menu)
- choicebox = select options from a given list
- enterbox = enter some text
- diropenbox/fileopenbox/filesavebox = GUI for file handling

Basic structure

```
 1  import easygui as eg
 2  import sys
 3
 4  while 1:
 5      #
 6      #your functions / boxes
 7      #
 8      msg = "do you want to continue?"
 9      title = "tool usage"
10      if eg.ccbox( msg, title ):
11          pass
12      else:
13          sys.exit( 0 )
```

Exercises G – Part1

- 1.1) Write a script to handle input of primer names and sequences! All information should be saved in a multiple FASTA file.

- 1.2) Write a script to return a matching primer sequence from a FASTA file based on a given primer name.

- 1.3) Write a script to combine both functionalities: return primer sequence, if name is already present OR generate new entry if primer name is novel.

Advanced Python

Overview

- Submission of processes to shell for execution and result handling
- BLAST result analysis
- Concept of plots via axes in matplotlib
- General figure types: plot, boxplot, barplot
  - Genome figures: scatter plots, box plots, chromosome plots, gene cluster plots
- Statistics in Python: tests and theoretical background
  - t-test, W-test, U-test, cor-test, X^2-test
- HTML construction

Analyze problem

- Split problem into parts => solutions for small parts are more generic and thus often available

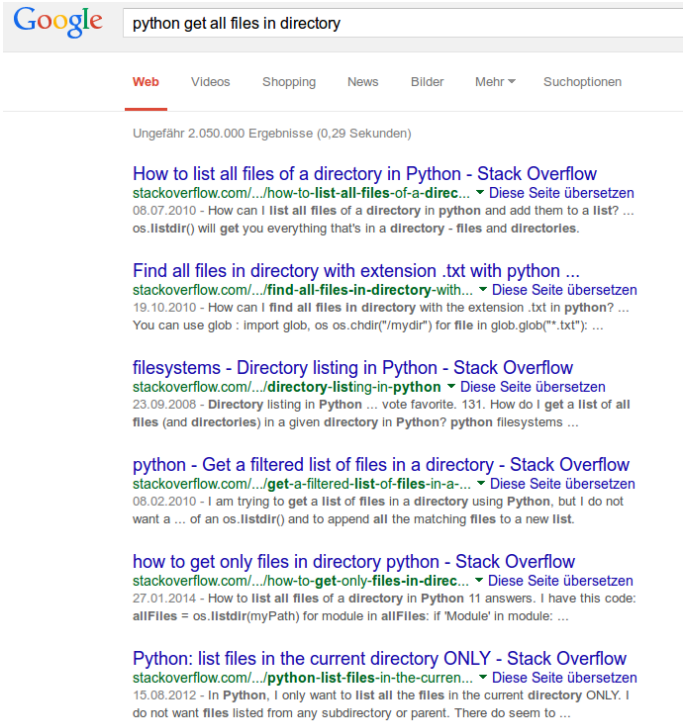
- Precise description of problem is necessary:

stackoverflow

- Best hit in approximately 90-95% of all cases
- Stackoverflow is chat forum for programmers (different languages)
- Problems are described in questions
- Answers often contain examples which can be directly applied
- Answers are judged by members => look for green marking and the number of positive votes

Example

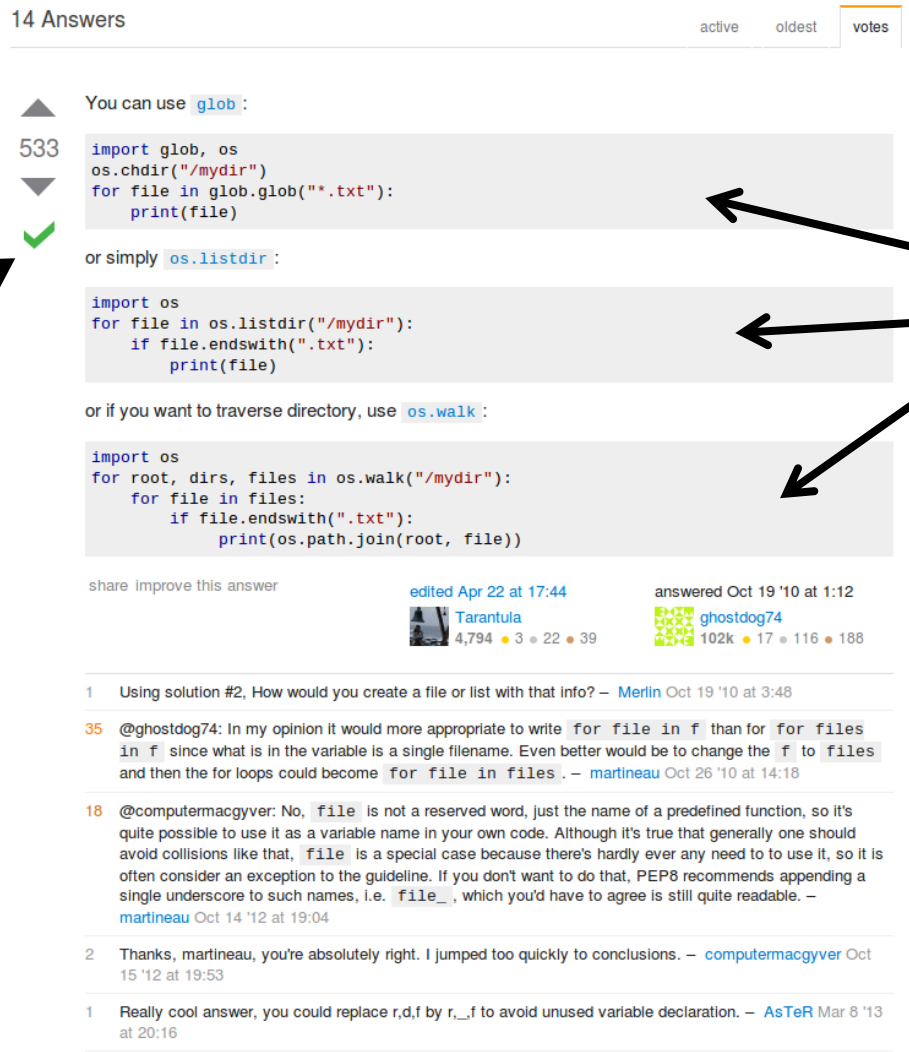
- Problem: get paths of all files in a certain directory
- Search expression: “python get all files in directory”



Example

533 user like this answer

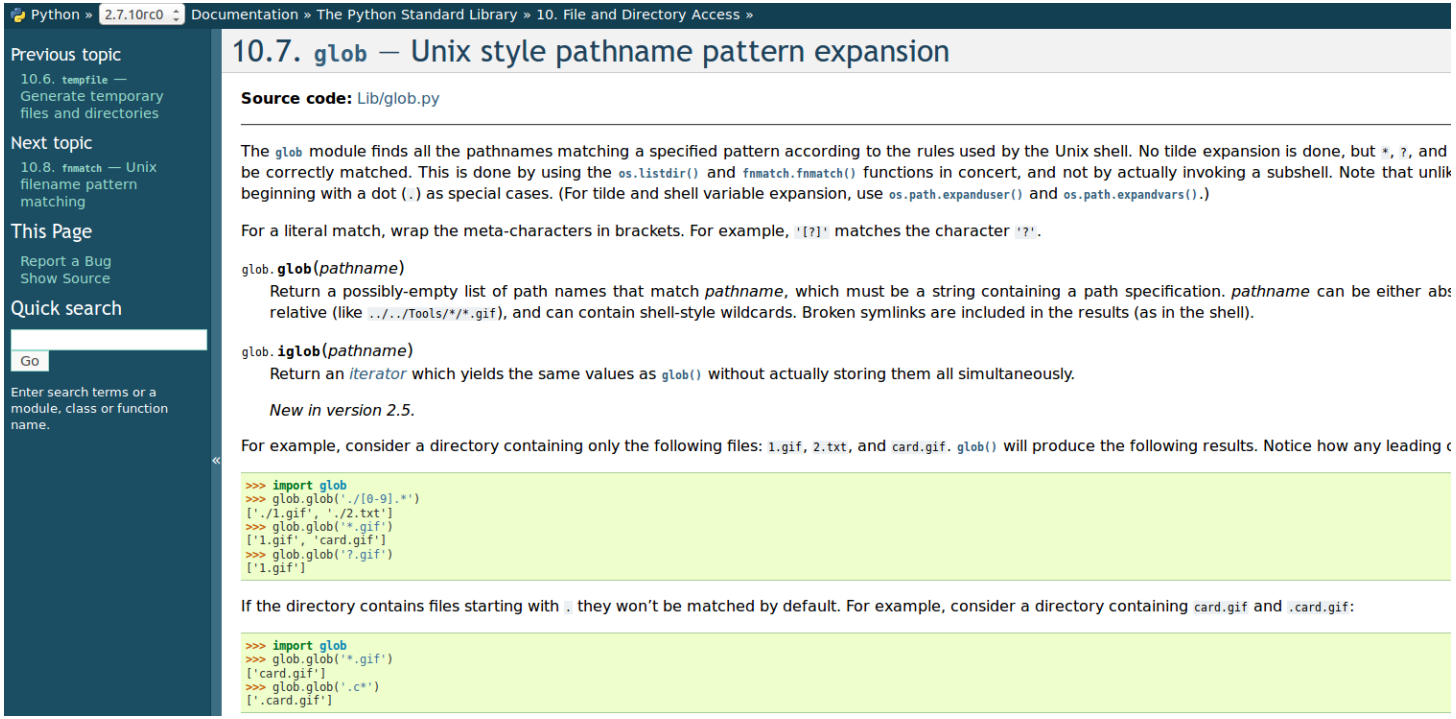
This answers solved the question



Three different ways to solve this problem are described

Official Python documentation

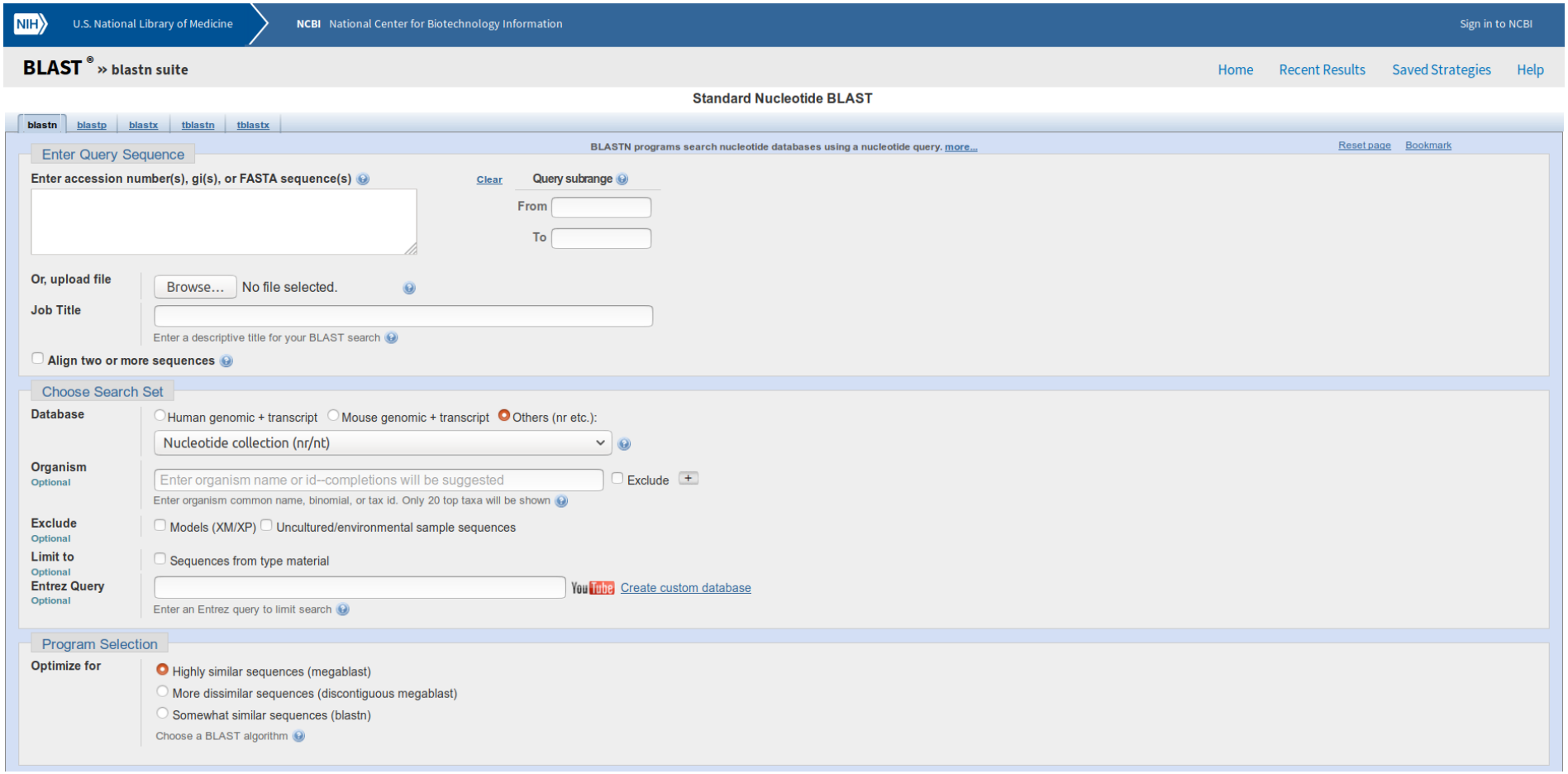
- Systematic documentation of all functions in a module with all possible arguments
- Sometimes examples are given as well



Books?

- There is enough out there ….
  - Google's Python Class: https://developers.google.com/edu/python/
  - Youtube: https://www.youtube.com/watch?v=tKTZoB2Vjuk

BLAST


U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Help
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm

How to execute processes via shell

- os.popen(command)

- Can be used to do almost everything via Python
- Python stops at line until command is completed
- Example:
  os.popen(‘mkdir my_new_popen_test‘)

Running BLAST

- BLAST = Basic Local Alignment Search Tool

  blastn –query <query_file> \
  -subject <subject_file> \ #-db <db_name>
  -out <output_file> \
  -outfmt 6
- Useful arguments:

  -evalue <FLOAT> #e-value cutoff
  -num_threads <INT> #number of parallel threads

BLAST results (outfmt 6)

1.	qseqid	query (e.g., gene) sequence id
2.	sseqid	subject (e.g., reference genome) sequence id
3.	pident	percentage of identical matches
4.	length	alignment length
5.	mismatch	number of mismatches
6.	gapopen	number of gap openings
7.	qstart	start of alignment in query
8.	qend	end of alignment in query
9.	sstart	start of alignment in subject
10.	send	end of alignment in subject
11.	evalue	expect value
12.	bitscore	bit score

(source: http://www.metagenomics.wiki/tools/blast/blastn-output-format-6)

BLAST result parsing

```python
def load_BLAST_results( input_file ):
    """! @brief load all BLAST results from file """

    data = []
    with open( input_file, "r" ) as f:
        line = f.readline()
        while line:
            parts = line.strip().split('\t')
            data.append( {  'query': parts[0],
                            'subject': parts[1],
                            'query_start': int( parts[6] ),
                            'query_end': int( parts[7] ),
                            'score': float( parts[-1] )
                        } )
            line = f.readline()
    return data
```

```
AT1G01010       NdCChr1.g1.t1   100.00  429     0       0       1       429     1       429     0.0      895
AT1G01010       NdCChr4.g18734.t1       32.84   469     247     15      1       428     2       443     4e-56    194
AT1G01010       NdCChr1.g127.t1 34.23   336     157     10      1       330     1       278     1e-41    152
AT1G01010       NdCChr1.g128.t1 32.38   349     157     14      1       331     1       288     4e-39    146
AT1G01010       NdCChr4.g18730.t1       39.33   178     95      5       1       175     2       169     1e-32    126
AT1G01010       NdCChr3.g12773.t1       39.63   164     89      2       1       162     1       156     1e-28    115
AT1G01010       NdCChr4.g22969.t1       40.74   162     79      5       5       159     11      162     3e-28    117
AT1G01010       NdCChr4.g18733.t1       40.00   165     90      5       1       162     2       160     4e-28    115
AT1G01010       NdCChr3.g17122.t1       42.31   156     74      6       5       153     15      161     4e-27    112
```

How to organize a Python script …

- Executable script under LINUX:

#!/usr/bin/env python #takes python version in path

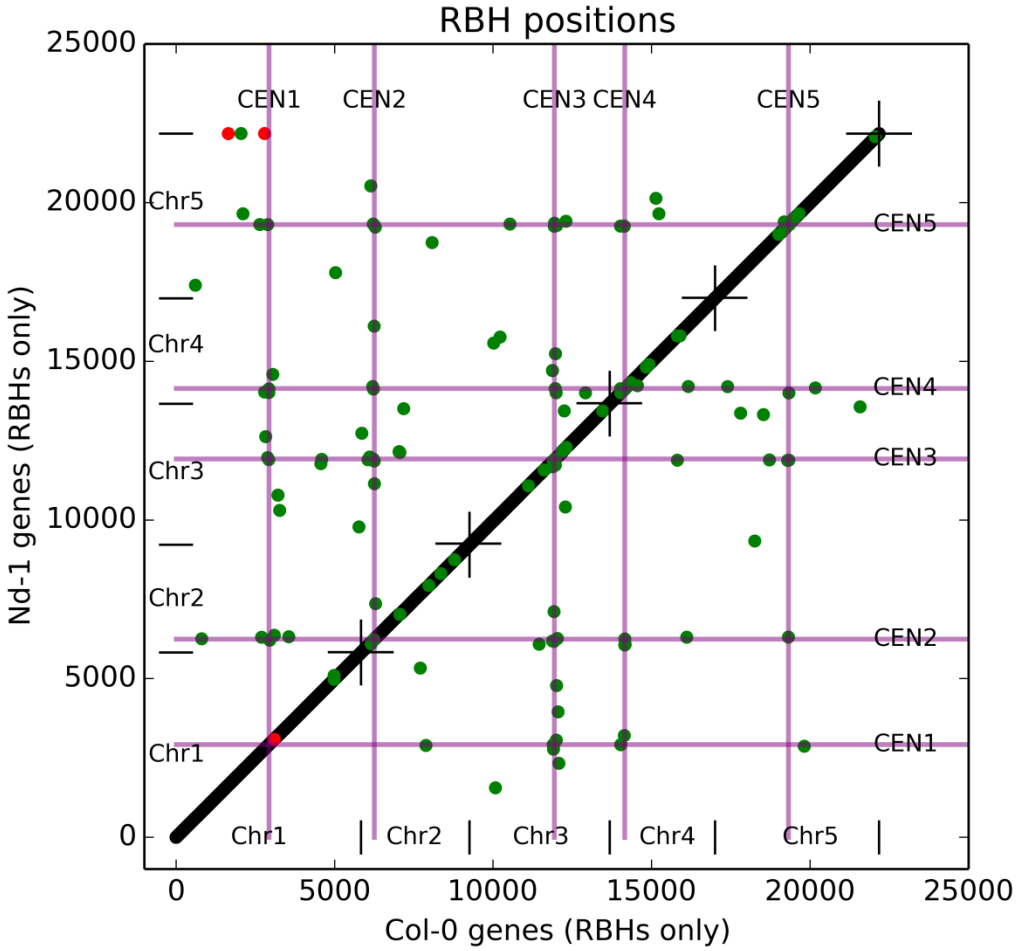
#!/usr/bin/python #specifies a specific python version

- Author
- Version
- Imports
- Usage

```
1  ### Boas Pucker ###
2  ### bpucker@cebitec.uni-bielefeld.de ###
3  ### v0.2 ###
4
5  __usage__ = """
6                  python construct_RNA_seq_coverage_file.py\n
7                  --in <BAM_FILE>
8                  --out <OUTPUT_FILE>
9
10                 --bam_is_sorted <PREVENTS_EXTRA_SORTING_OF_BAM_FILE>
11
12                 feature requests and bug reports: bpucker@cebitec.uni-bielefeld.de
13                 """
14
15 __cite__ = """ Pucker & Brockington, 2018: https://doi.org/10.1186/s12864-018-5360-z """
16
17
18 import os, sys
19
20 # --- end of imports --- #
21
22 def main( arguments ):
```

matplotlib

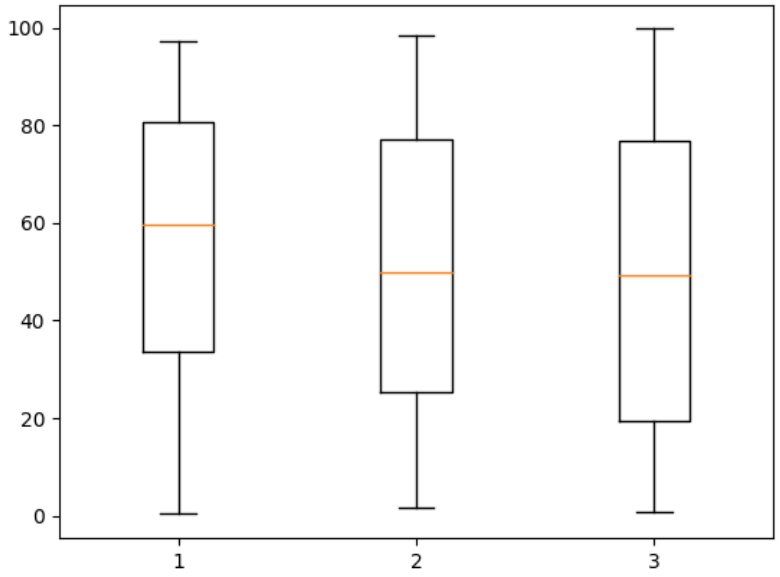
- „import matplotlib.pyplot as plt“
- Visualization of complex data
- Automatic generation of plots
- Amazing customization options



(Pucker *et al*., 2016)

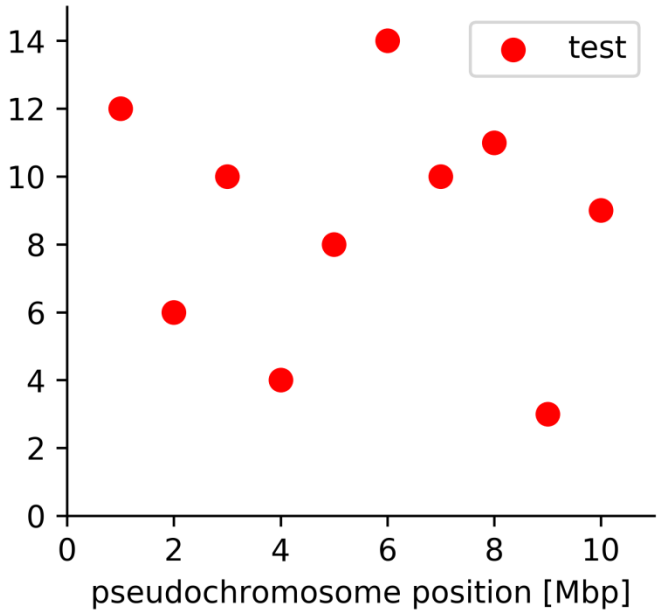
Box plot

```
1   import matplotlib.pyplot as plt
2   import numpy as np
3
4   d1 = np.random.rand(50) * 100   #generate random numbers
5   d2 = np.random.rand(50) * 100
6   d3 = np.random.rand(50) * 100
7
8   data = [d1, d2, d3] # multiple box plots on one figure
9
10  plt.figure()
11  plt.boxplot(data)
12  plt.show()
```



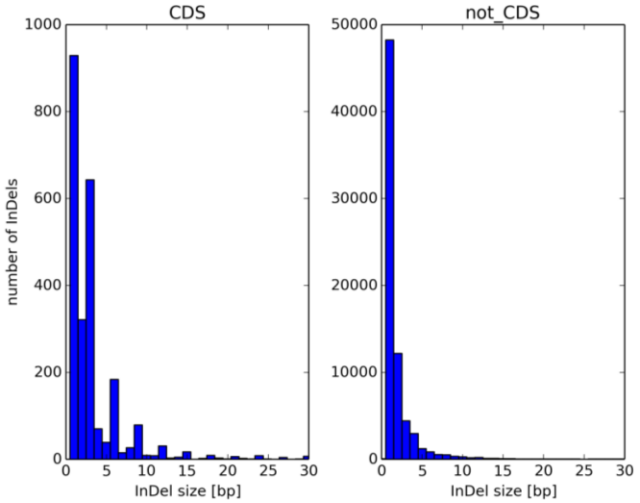
Scatter plot

```python
import matplotlib.pyplot as plt

fig, ax = plt.subplots( figsize=( 10, 4 ) )   #defining size of plot

x_values = [ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ]
y_values = [ 12, 6, 10, 4, 8, 15, 10, 11, 3, 9 ]

ax.scatter( x_values, y_values, color="red", s=10, marker="o", label="test" )
#setting color, marker size, marker shape and label of this group

ax.legend( numpoints=1 )
#each group is represented by only one marker in the legend (default=3)

ax.set_xlim( 0, 11 )    #set range of x-axis
ax.set_ylim( 0, 15 )    #set range of y-axis

ax.set_xlabel( "pseudochromosome position [Mbp]" )

ax.spines["top"].set_visible(False)      #remove lines and ticks
ax.spines["right"].set_visible(False)    #remove lines and ticks

plt.subplots_adjust(left=0.05, right=0.99, top=0.97, bottom=0.12)
#adjust size of plot within figure

plt.show()
fig.savefig( "my_plot.png", dpi=600 )   #write figure into output file
plt.close( "all" )   #destroy created figures (cleaning up)
```



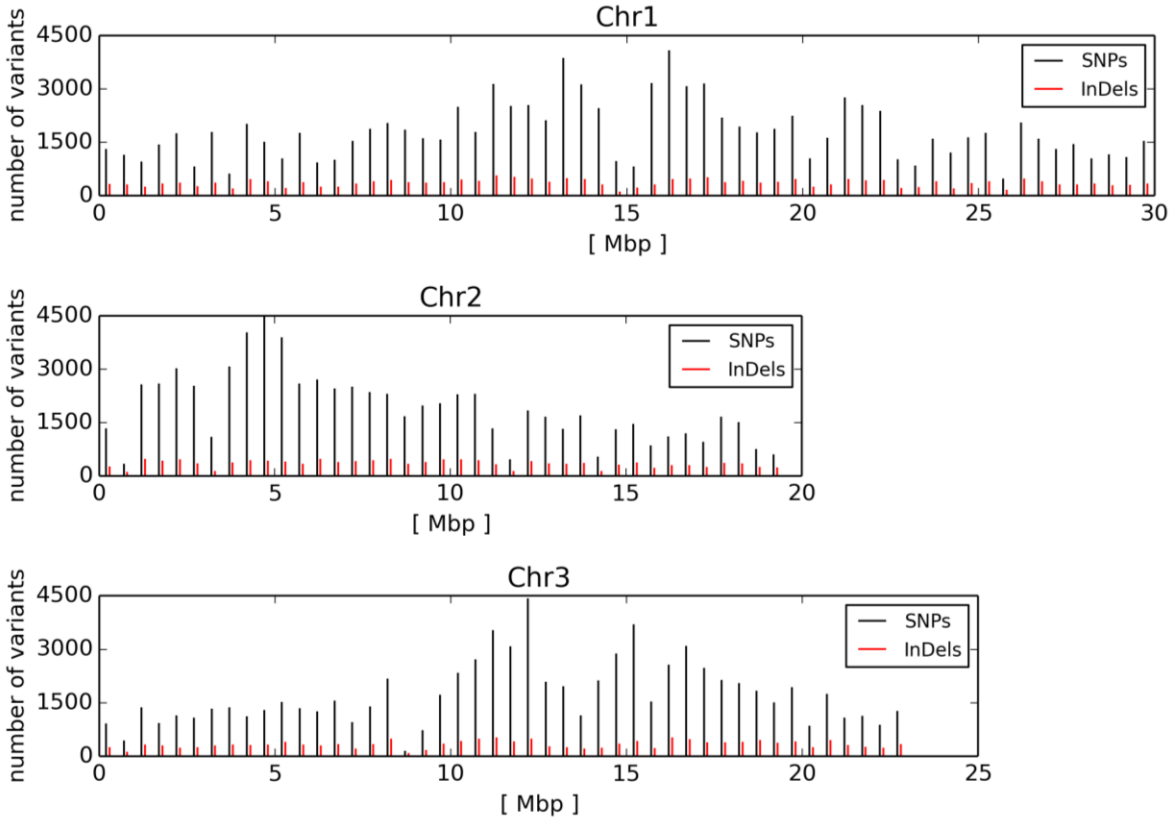
Histogram

```python
import matplotlib.pyplot as plt

# --- end of imports --- #

gene_space = [  3, 3, 6, 6, 9, 9, 12, 3, 3, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
                11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
                12, 15, 18, 21, 24, 27, 30 ]
intergenic = [ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9,
                1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 1, 2, 1 ]


fig, ( ax1, ax2 ) = plt.subplots( 1, 2, sharey=False)
counts, bins, patches = ax1.hist( gene_space, bins=max( gene_space ), align="left" )
ax1.set_title( "CDS" )
ax1.set_xlim( 0, 30 )
ax1.set_xlabel( "InDel size [bp]" )
ax1.set_ylabel( "number of InDels" )

counts, bins, patches = ax2.hist( intergenic, bins=max( intergenic ), align="left" )
ax2.set_title( "not_CDS" )
ax2.set_xlim( 0, 30 )
ax2.set_xlabel( "InDel size [bp]" )
plt.subplots_adjust( wspace=0.3 )   #increase space between figures

plt.show()
fig.savefig( prefix + "InDel_size_distribution.png", dpi=300 )
plt.close('all')
```



(Pucker et al., 2016)

‘barplot‘ figure

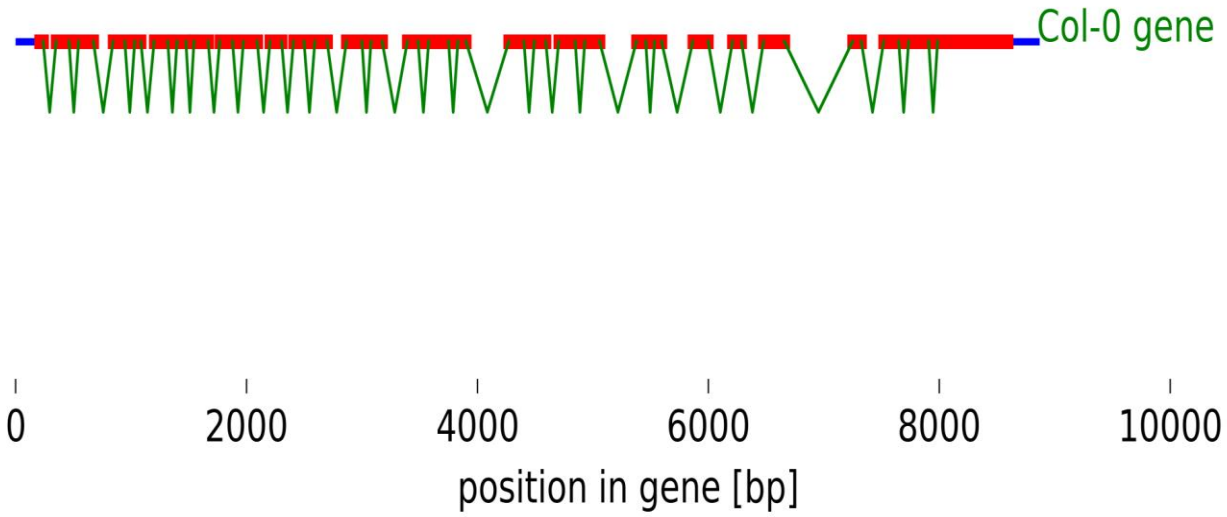


barplots.py generates barplots at specific positions by drawing a normal line

(script is available in course repository)

(Pucker et al., 2016)

Gene structure plot



(Pucker et al., 2017)

gene_structure_plot.py generates visualizations of gene/transcript structures based on GFF annotations

Exercise H: operon structure plot

- Construct a figure to illustrate the order and orientation of genes in the gum gene cluster in *Xanthomonas campestris* pv. campestris!

Statistics

Theoretical background

- Normal distribution plot

Compare observed sample against the expected distribution

H0 = sample was taken from distribution

H0 can only be rejected or kept due to insufficient evidence against it

H0 can NEVER be confirmed!

Shapiro-Wilk test

scipy.stats.shapiro

`scipy.stats.` **`shapiro`**(*x*, *a=None*, *reta=False*) [source]

Perform the Shapiro-Wilk test for normality.

The Shapiro-Wilk test tests the null hypothesis that the data was drawn from a normal distribution.

Parameters:

x : *array_like*
Array of sample data.

a : *array_like, optional*
Array of internal parameters used in the calculation. If these are not given, they will be computed internally. If x has length n, then a must have length n/2.

reta : *bool, optional*
Whether or not to return the internally computed a values. The default is False.

Returns:

W : *float*
The test statistic.

p-value : *float*
The p-value for the hypothesis test.

a : *array_like, optional*
If *reta* is True, then these are the internally computed "a" values that may be passed into this function on future calls.

See also:

`anderson` The Anderson-Darling test for normality

`kstest` The Kolmogorov-Smirnov test for goodness of fit.

Notes

The algorithm used is described in [R634] but censoring parameters as described are not implemented. For N > 5000 the W test statistic is accurate but the p-value may not be.

The chance of rejecting the null hypothesis when it is true is close to 5% regardless of sample size.

https://docs.scipy.org

Correlation

scipy.stats.pearsonr

scipy.stats.pearsonr(*x*, *y*) [source]

Calculates a Pearson correlation coefficient and the p-value for testing non-correlation.

The Pearson correlation coefficient measures the linear relationship between two datasets. Strictly speaking, Pearson's correlation requires that each dataset be normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so.

Parameters: **x** : *(N,) array_like*
Input
y : *(N,) array_like*
Input

Returns: *(Pearson's correlation coefficient,*
2-tailed p-value)

https://docs.scipy.org

t-test

scipy.stats.ttest_ind

scipy.stats.**ttest_ind**(*a*, *b*, *axis=0*, *equal_var=True*, *nan_policy='propagate'*) [source]

Calculates the T-test for the means of *two independent* samples of scores.

This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances by default.

Parameters:

a, b : *array_like*
The arrays must have the same shape, except in the dimension corresponding to *axis* (the first, by default).

axis : *int or None, optional*
Axis along which to compute test. If None, compute over the whole arrays, *a*, and *b*.

equal_var : *bool, optional*
If True (default), perform a standard independent 2 sample test that assumes equal population variances [R643]. If False, perform Welch's t-test, which does not assume equal population variance [R644].
New in version 0.11.0.

nan_policy : *{'propagate', 'raise', 'omit'}, optional*
Defines how to handle when input contains nan. 'propagate' returns nan, 'raise' throws an error, 'omit' performs the calculations ignoring nan values. Default is 'propagate'.

Returns:

statistic : *float or array*
The calculated t-statistic.

pvalue : *float or array*
The two-tailed p-value.

Notes

We can use this test, if we observe two independent samples from the same or different population, e.g. exam scores of boys and girls or of two ethnic groups. The test measures whether the average (expected) value differs significantly across samples. If we observe a large p-value, for example larger than 0.05 or 0.1, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, e.g. 1%, 5% or 10%, then we reject the null hypothesis of equal averages.

https://docs.scipy.org

Wilcoxon test

scipy.stats.wilcoxon

scipy.stats.wilcoxon(*x*, *y=None*, *zero_method='wilcox'*, *correction=False*) [source]

Calculate the Wilcoxon signed-rank test.

The Wilcoxon signed-rank test tests the null hypothesis that two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences x - y is symmetric about zero. It is a non-parametric version of the paired T-test.

Parameters:

x : *array_like*
The first set of measurements.

y : *array_like, optional*
The second set of measurements. If *y* is not given, then the *x* array is considered to be the differences between the two sets of measurements.

zero_method : *string, {"pratt", "wilcox", "zsplit"}, optional*

"pratt":
Pratt treatment: includes zero-differences in the ranking process (more conservative)

"wilcox":
Wilcox treatment: discards all zero-differences

"zsplit":
Zero rank split: just like Pratt, but spliting the zero rank between positive and negative ones

correction : *bool, optional*
If True, apply continuity correction by adjusting the Wilcoxon rank statistic by 0.5 towards the mean value when computing the z-statistic. Default is False.

Returns:

T : *float*
The sum of the ranks of the differences above or below zero, whichever is smaller.

p-value : *float*
The two-sided p-value for the test.

Notes

Because the normal approximation is used for the calculations, the samples used should be large. A typical rule is to require that n > 20.

Mann-Whitney rank test

scipy.stats.mannwhitneyu

`scipy.stats.` **`mannwhitneyu`**(*x*, *y*, *use_continuity=True*, *alternative=None*) [source]

Computes the Mann-Whitney rank test on samples x and y.

Parameters:

x, y : *array_like*
Array of samples, should be one-dimensional.

use_continuity : *bool, optional*
Whether a continuity correction (1/2.) should be taken into account. Default is True.

alternative : *None (deprecated), 'less', 'two-sided', or 'greater'*
Whether to get the p-value for the one-sided hypothesis ('less' or 'greater') or for the two-sided hypothesis ('two-sided'). Defaults to None, which results in a p-value half the size of the 'two-sided' p-value and a different U statistic. The default behavior is not the same as using 'less' or 'greater': it only exists for backward compatibility and is deprecated.

Returns:

statistic : *float*
The Mann-Whitney U statistic, equal to min(U for x, U for y) if *alternative* is equal to None (deprecated; exists for backward compatibility), and U for y otherwise.

pvalue : *float*
p-value assuming an asymptotic normal distribution. One-sided or two-sided, depending on the choice of *alternative*.

Notes

Use only when the number of observation in each sample is > 20 and you have 2 independent samples of ranks. Mann-Whitney U is significant if the u-obtained is LESS THAN or equal to the critical value of U.

This test corrects for ties and by default uses a continuity correction.

https://docs.scipy.org

Chi square test

scipy.stats.chisquare

scipy.stats.**chisquare**(*f_obs*, *f_exp=None*, *ddof=0*, *axis=0*) [source]

Calculates a one-way chi square test.

The chi square test tests the null hypothesis that the categorical data has the given frequencies.

Parameters:

f_obs : *array_like*
Observed frequencies in each category.

f_exp : *array_like, optional*
Expected frequencies in each category. By default the categories are assumed to be equally likely.

ddof : *int, optional*
"Delta degrees of freedom": adjustment to the degrees of freedom for the p-value. The p-value is computed using a chi-squared distribution with `k - 1 - ddof` degrees of freedom, where *k* is the number of observed frequencies. The default value of *ddof* is 0.

axis : *int or None, optional*
The axis of the broadcast result of *f_obs* and *f_exp* along which to apply the test. If axis is None, all values in *f_obs* are treated as a single data set. Default is 0.

Returns:

chisq : *float or ndarray*
The chi-squared test statistic. The value is a float if *axis* is None or *f_obs* and *f_exp* are 1-D.

p : *float or ndarray*
The p-value of the test. The value is a float if *ddof* and the return value *chisq* are scalars.

See also:

`power_divergence`, `mstats.chisquare`

Notes

This test is invalid when the observed or expected frequencies in each category are too small. A typical rule is that all of the observed and expected frequencies should be at least 5.

The default degrees of freedom, k-1, are for the case when no parameters of the distribution are estimated. If p parameters are estimated by efficient maximum likelihood then the correct degrees of freedom are k-1-p. If the parameters are estimated in a different way, then the dof can be between k-1-p and k-1. However, it is also possible that the asymptotic distribution is not a chisquare, in which case this test is not appropriate.

https://docs.scipy.org

Exercise I: analyze the unknown data

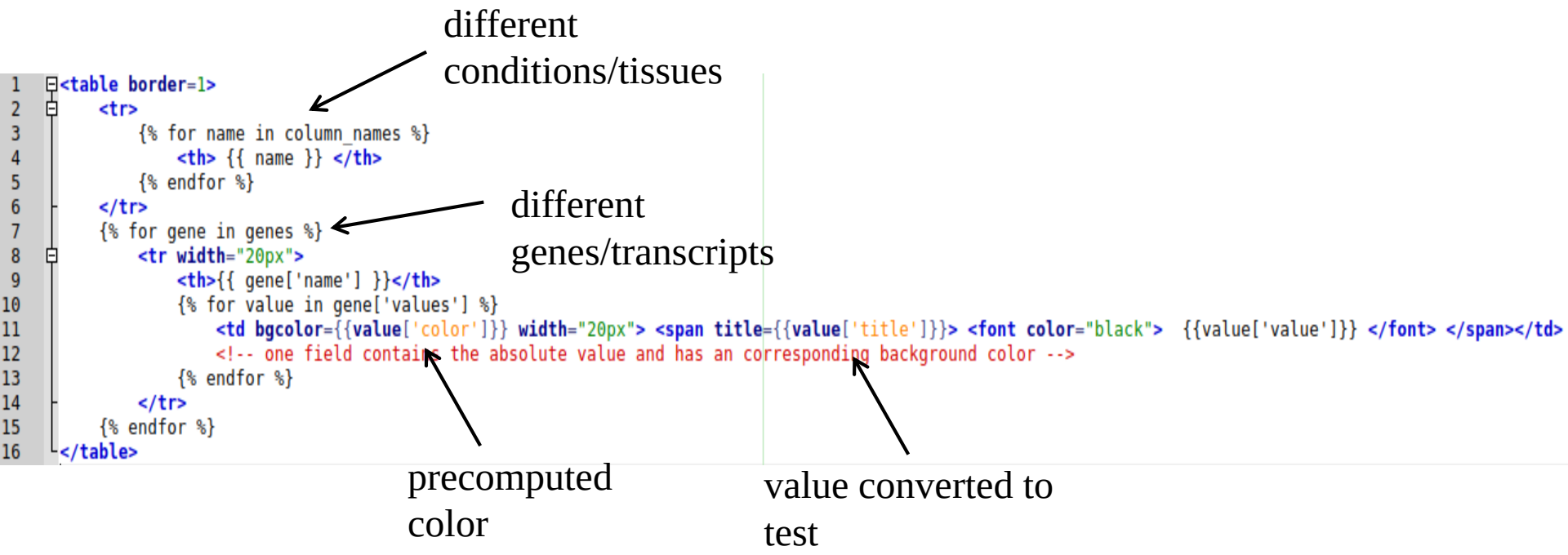
- Construct a suitable visualization!
- Analyze distribution and trends!
- Apply statistical test to investigate difference!

HTML

- Construction of a HTML-based heatmap

gene ID	0H	4H	Salt	Heat	Inflorescence	leaf	Root	seedlings_HiK	seedlings_HiK2	Leaf_SSC
(p	0.02	0.0	0.0	0.0	0.0	0.0	4.33	1.44	1.2	0.0
(c	0.0	0.0	0.0	0.0	0.0	0.0	1.17	0.11	0.0	0.0
(c	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.11	0.4	0.0
p	1.76	29.23	14.65	6.11	203.75	265.2	0.83	199.33	199.7	54.4
o	0.0	0.0	0.0	0.01	1.38	0.0	0.0	0.0	0.1	0.0
(	0.0	0.03	0.01	0.0	0.25	0.0	0.0	0.0	0.1	0.02
y	0.01	0.01	0.0	0.06	13.88	0.0	0.0	0.11	0.1	0.01
u	6.25	58.89	41.42	15.61	69.5	5.4	3.83	157.22	163.1	22.49
h	0.86	5.04	0.09	0.24	33.5	0.8	12.67	8.78	8.5	1.07
y	10.74	43.0	4.78	6.45	29.0	1.2	7.17	62.67	67.0	25.48
e	19.29	14.22	6.14	5.21	4.88	1.6	11.5	28.11	24.2	13.68
w	35.33	45.16	52.79	70.1	17.5	74.2	6.67	21.89	22.6	65.95
q	6.03	24.4	2.69	2.38	80.0	35.6	19.83	55.22	59.9	11.16
k	5.74	15.9	1.35	4.56	57.5	3.0	1.17	9.44	7.5	8.47
s	0.27	1.74	0.07	0.25	11.38	0.2	0.0	0.78	0.4	0.11
i	0.04	0.0	0.0	0.0	36.0	0.0	0.0	0.89	1.1	0.0
[illegible]	0.0	0.0	0.0	0.0	18.0	0.0	0.0	0.67	0.7	0.0
[illegible]	9.84	14.92	17.27	7.68	16.13	27.0	26.0	26.89	23.7	11.58
ev	17.99	9.41	7.94	8.45	11.38	21.2	13.83	19.22	19.7	14.81

HTML template




```
<table border=1>
    <tr>
        {% for name in column_names %}
            <th> {{ name }} </th>
        {% endfor %}
    </tr>
    {% for gene in genes %}
        <tr width="20px">
            <th>{{ gene['name'] }}</th>
            {% for value in gene['values'] %}
                <td bgcolor={{value['color']}} width="20px"> <span title={{value['title']}}> <font color="black">  {{value['value']}} </font> </span></td>
                <!-- one field contains the absolute value and has an corresponding background color -->
            {% endfor %}
        </tr>
    {% endfor %}
</table>
```

construct_heatmap.py reads values from text file and prepares data structures to fill this template

HTML document can be converted to PDF

Exercise J: construct heatmap

- Read data table and construct heatmap for the gene expression in HTML!
- Add mouse-over effect to display functional gene annotation!

Biological background of presented examples

- Nd-1 genome assembly (Pucker *et al.*, 2016):
  - https://doi.org/10.1371/journal.pone.0164321
- Non-canonical splice sites (Pucker *et al.*, 2017):
  - https://doi.org/10.1186/s13104-017-2985-y
- *Croton tiglium* transcriptome assembly (Haak *et al.*, 2018):
  - https://doi.org/10.3389/fmolb.2018.00062
- Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes (Pucker & Brockington, 2018):
  - https://doi.org/10.1186/s12864-018-5360-z
- A Chromosome-level Sequence Assembly Reveals the Structure of the *Arabidopsis thaliana* Nd-1 Genome and ist Gene Set (Pucker *et al.*, 2019):
  - https://doi.org/10.1101/407627
- NAVIP: https://github.com/bpucker/NAVIP