python™

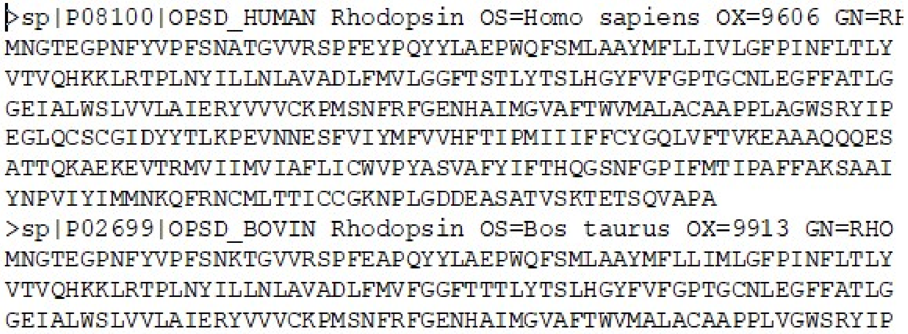
MySQL®

ISC

NAME: ADRIÁN GARCÍA RECIO
BIOINFORMATICS: FINAL PROJECT
DATE: 13/05/2018

IDENTITY AND SIMILARITY CALCULATOR

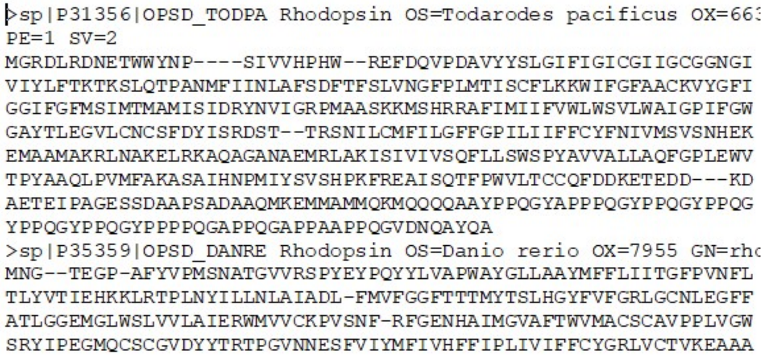
- **BIOLOGIC PROBLEM:**
  - Multiple Sequence Alignment



```
>sp|P08100|OPSD_HUMAN Rhodopsin OS=Homo sapiens OX=9606 GN=RH
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLAGWSRYIP
EGLQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIIIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSNFGPIFMTIPAFFAKSAAI
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
>sp|P02699|OPSD_BOVIN Rhodopsin OS=Bos taurus OX=9913 GN=RHO
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
```

seqs.fasta

→



```
>sp|P31356|OPSD_TODPA Rhodopsin OS=Todarodes pacificus OX=663
PE=1 SV=2
MGRDLRDNETWWYNP----SIVVHPHW--REFDQVPDAVYYSLGIFIGICGIIGCGGNGI
VIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCFLKKWIFGFAACKVYGFI
GGIFGFMSIMTMAMISIDRYNVIGRPMAASKKMSHRRAFIMIIFVWLWSVLWAIGPIFGW
GAYTLEGVLCNCSFDYISRDST--TRSNILCMFILGFFGPILIIFFCYFNIVMSVSNHEK
EMAAMAKRLNAKELRKAQAGANAEMRLAKISIVIVSQFLLSWSPYAVVALLAQFGPLEWV
TPYAAQLPVMFAKASAIHNPMIYSVSHPKFREAISQTFPWVLTCCQFDDKETEDD---KD
AETEIPAGESSDAAPSADAAQMKEMMAMMQKMQQQQAAYPPQGYAPPPQGYPPQGYPPQG
YPPQGYPPQGYPPPPQGAPPQGAPPAAPPQGVDNQAYQA
>sp|P35359|OPSD_DANRE Rhodopsin OS=Danio rerio OX=7955 GN=rh
MNG--TEGP-AFYVPMSNATGVVRSPYEYPQYYLVAPWAYGLLAAYMFFLIITGFPVNFL
TLYVTIEHKKLRTPLNYILLNLAIADL-FMVFGGFTTTMYTSLHGYFVFGRLGCNLEGFF
ATLGGEMGLWSLVVLAIERWMVVCKPVSNF-RFGENHAIMGVAFTWVMACSCAVPPLVGW
SRYIPEGMQCSCGVDYYTRTPGVNNESFVIYMFIVHFFIPLIVIFFCYGRLVCTVKEAAA
```

msa.fasta

- **WHY I LIKE RESOLD THIS PROBLEM?**

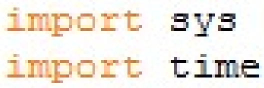
MASTER →

ALIGNMENT
ALIGNMENT
ALIGNMENT
ALIGNMENT
ALIGNMENT
ALIGNMENT

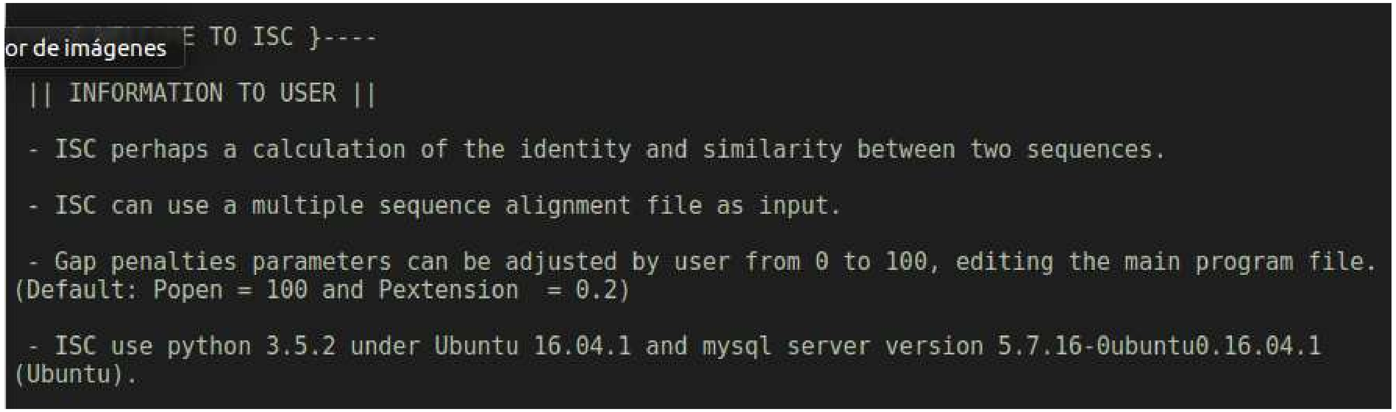
→

SCHEME OF PROGRAM STRUCTURE

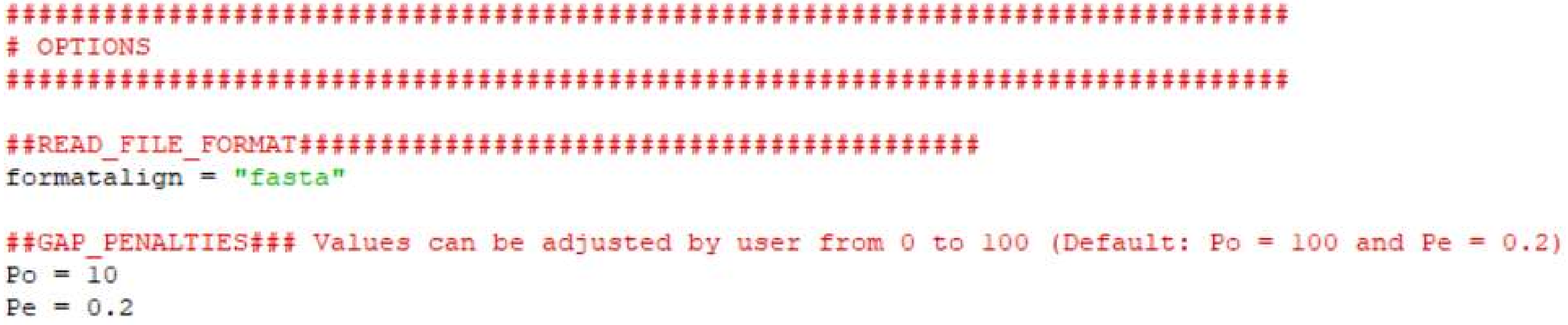
- **1- INTRODUCTION:**



```
import sys
import time
```



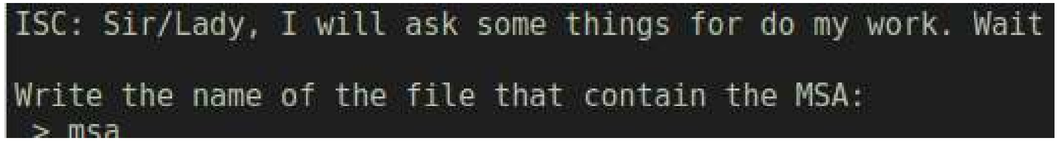
```
E TO ISC }----

 || INFORMATION TO USER ||

 - ISC perhaps a calculation of the identity and similarity between two sequences.

 - ISC can use a multiple sequence alignment file as input.

 - Gap penalties parameters can be adjusted by user from 0 to 100, editing the main program file.
(Default: Popen = 100 and Pextension  = 0.2)

 - ISC use python 3.5.2 under Ubuntu 16.04.1 and mysql server version 5.7.16-0ubuntu0.16.04.1
(Ubuntu).
```



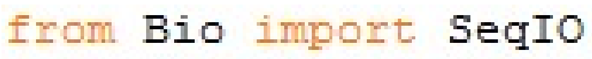
```
################################################################################
# OPTIONS
################################################################################

##READ_FILE_FORMAT###########################################
formatalign = "fasta"

##GAP_PENALTIES### Values can be adjusted by user from 0 to 100 (Default: Po = 100 and Pe = 0.2)
Po = 10
Pe = 0.2
```

SCHEME OF PROGRAM STRUCTURE

- **2- READ THE MSA FILE:**

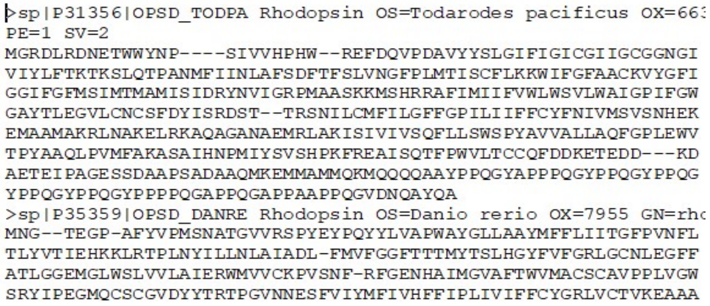


```
ISC: Sir/Lady, I will ask some things for do my work. Wait

Write the name of the file that contain the MSA:
 > msa
```

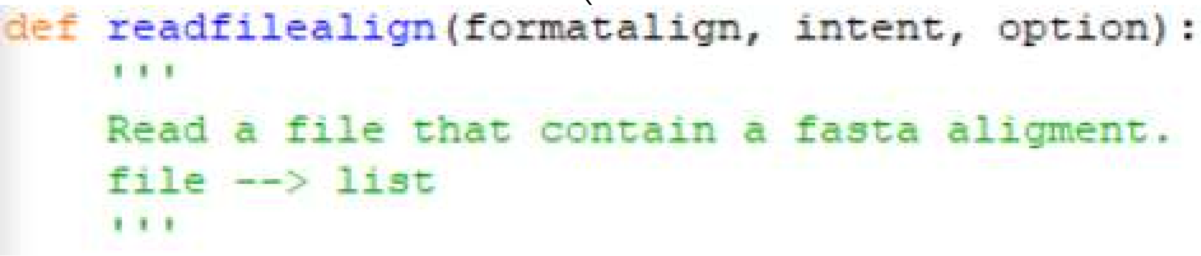


```
from Bio import SeqIO
```

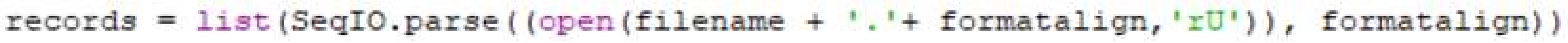
msa.fasta →



```
>sp|P31356|OPSD_TODPA Rhodopsin OS=Todarodes pacificus OX=663
PE=1 SV=2
MGRDLRDNETWWYNP----SIVVHPHW--REFDQVPDAVYYSLGIFIGICGIIGCGGNGI
VIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCFLKKWIFGFAACKVYGFI
GGIFGFMSIMTMAMISIDRYNVIGRPMAASKKMSHRRAFIMIIFVWLWSVLWAIGPIFGW
GAYTLEGVLCNCSFDYISRDST--TRSNILCMFILGFFGPILIIFFCYFNIVMSVSNHEK
EMAAMAKRLNAKELRKAQAGANAEMRLAKISIVIVSQFLLSWSPYAVVALLAQFGPLEWV
TPYAAQLPVMFAKASAIHNPMIYSVSHPKFREAISQTFPWVLTCCQFDDKETEDD---KD
AETEIPAGESSDAAPSADAAQMKEMMAMMQKMQQQQAAYPPQGYAPPPQGYPPQGYPPQG
YPPQGYPPQGYPPPPQGAPPQGAPPAAPPQGVDNQAYQA
>sp|P35359|OPSD_DANRE Rhodopsin OS=Danio rerio OX=7955 GN=rho
MNG--TEGP-AFYVPMSNATGVVRSPYEYPQYYLVAPWAYGLLAAYMFFLIITGFPVNFL
TLYVTIEHKKLRTPLNYILLNLAIADL-FMVFGGFTTTMYTSLHGYFVFGRLGCNLEGFF
ATLGGEMGLWSLVVLAIERWMVVCKPVSNF-RFGENHAIMGVAFTWVMACSCAVPPLVGW
SRYIPEGMQCSCGVDYYTRTPGVNNESFVIYMFIVHFFIPLIVIFFCYGRLVCTVKEAAA
```



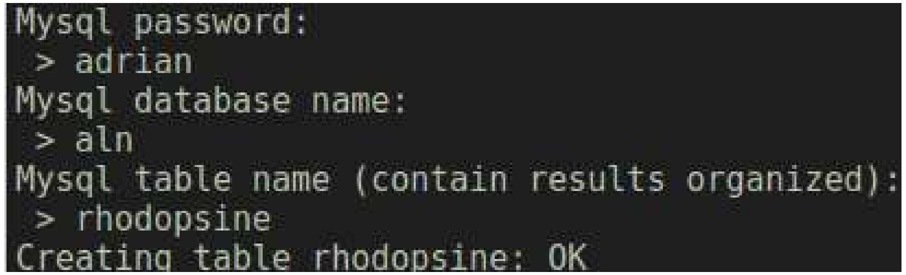
```
def readfilealign(formatalign, intent, option):
    '''
    Read a file that contain a fasta aligment.
    file --> list
    '''
```



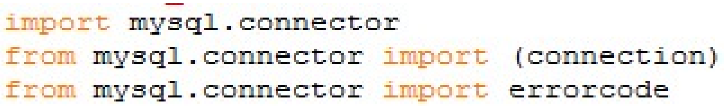
```
records = list(SeqIO.parse((open(filename + '.'+ formatalign,'rU')), formatalign))
```

SCHEME OF PROGRAM STRUCTURE

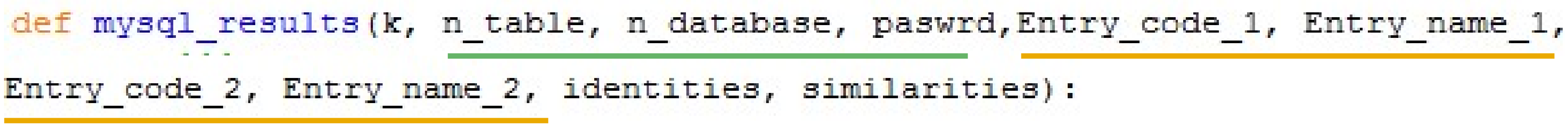
- 3- CREATE THE DATABASE MYSQL:



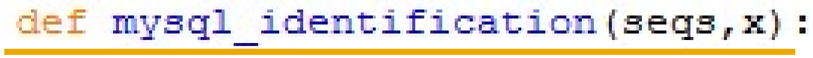
```
Mysql password:
 > adrian
Mysql database name:
 > aln
Mysql table name (contain results organized):
 > rhodopsine
Creating table rhodopsine: OK
```



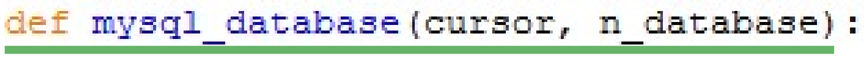
```
import mysql.connector
from mysql.connector import (connection)
from mysql.connector import errorcode
```



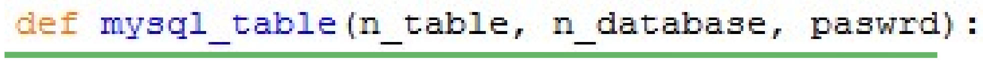
```
def mysql_results(k, n_table, n_database, paswrd,Entry_code_1, Entry_name_1,
Entry_code_2, Entry_name_2, identities, similarities):
```



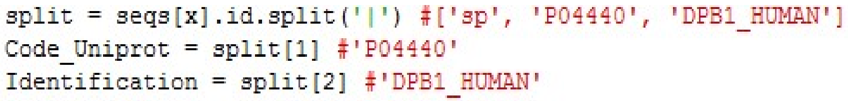
```
def mysql_identification(seqs,x):
```



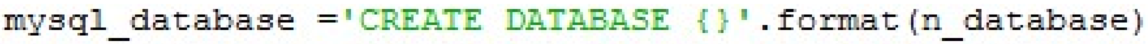
```
def mysql_database(cursor, n_database):
```



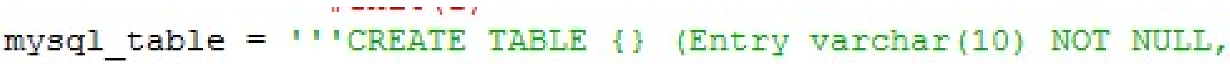
```
def mysql_table(n_table, n_database, paswrd):
```



```
split = seqs[x].id.split('|') #['sp', 'P04440', 'DPB1_HUMAN']
Code_Uniprot = split[1] #'P04440'
Identification = split[2] #'DPB1_HUMAN'
```



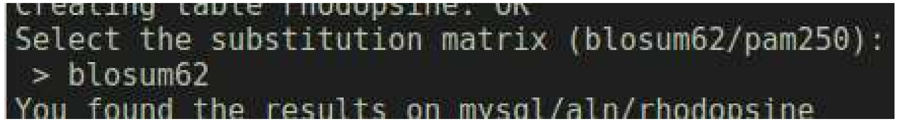
```
mysql_database ='CREATE DATABASE {}'.format(n_database)
```



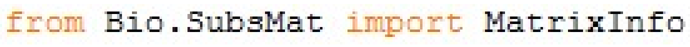
```
mysql_table = '''CREATE TABLE {} (Entry varchar(10) NOT NULL,
```

SCHEME OF PROGRAM STRUCTURE

- ## 4- SELECT SUBSTITUTION MATRIX:

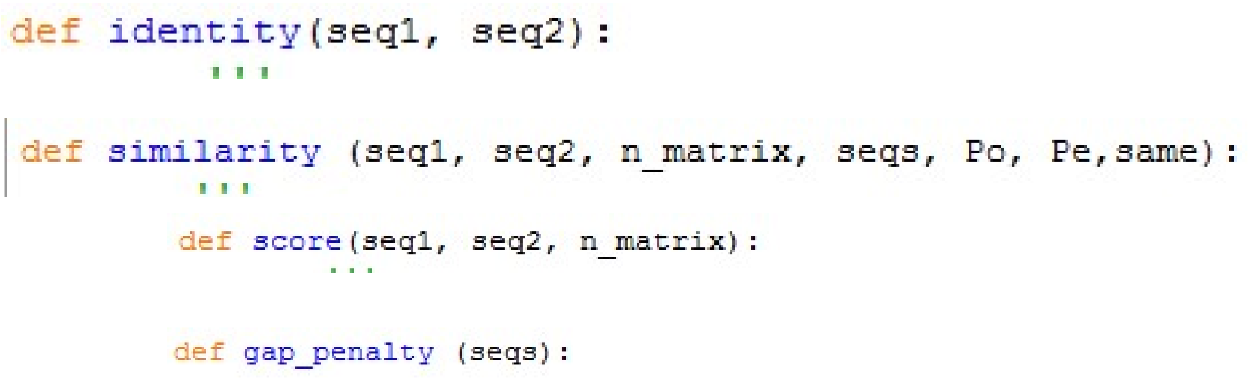


```
Creating table rhodopsine: OK
Select the substitution matrix (blosum62/pam250):
 > blosum62
You found the results on mysql/aln/rhodopsine
```



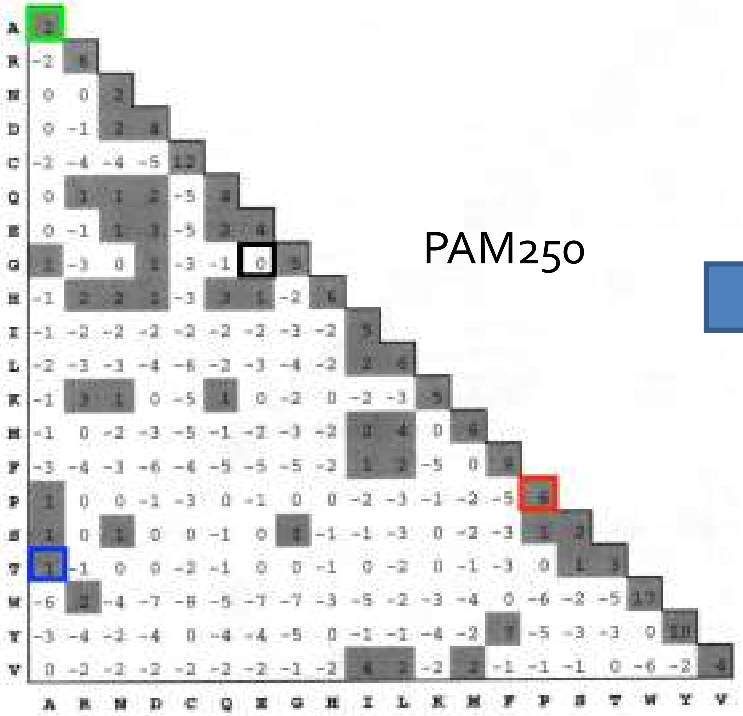
```
from Bio.SubsMat import MatrixInfo
```

- ## 5- CALCULATE IDENTITY AND SIMILARITY:



```
def identity(seq1, seq2):
        '''

def similarity (seq1, seq2, n_matrix, seqs, Po, Pe,same):
       '''
        def score(seq1, seq2, n_matrix):
              ...

        def gap_penalty (seqs):
```

SCHEME OF PROGRAM STRUCTURE

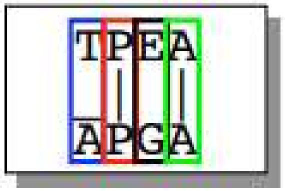


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

PAM250

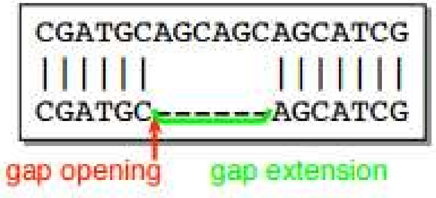
IDENTITY:

$$ID|SIM_{\%} = 100 * \frac{Identical\,|\,Similar\;Residues}{Sequence\;Length}$$



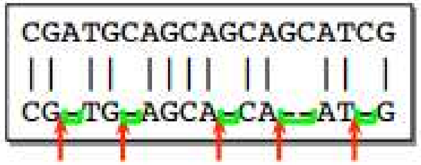
SIMILARITY:

$$S = ((\sum M_{ij}) + oP_o + eP_e)/\sum M_{ii}$$



Po = 10

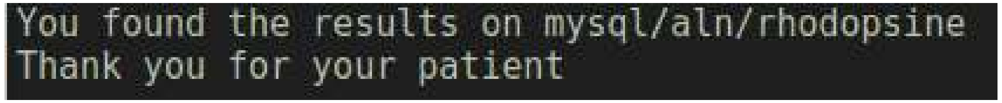
Pe= 0,2



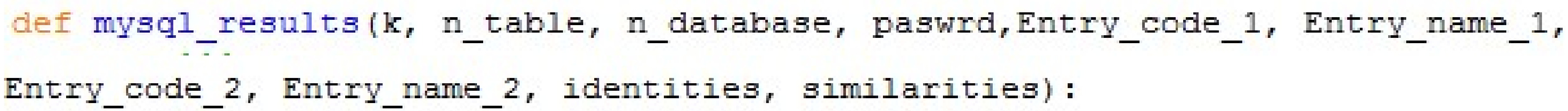
Po = 10 Po = 10 Po = 10 Po = 10 Po = 10

SCHEME OF PROGRAM STRUCTURE

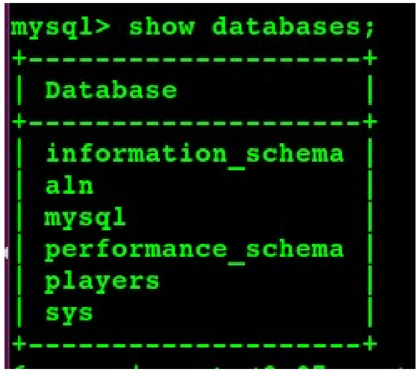
- ## 6- SAVE RESULTS ON MYSQL:



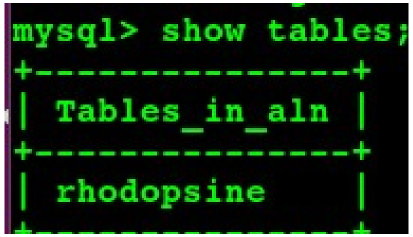
```
You found the results on mysql/aln/rhodopsine
Thank you for your patient
```



```
def mysql_results(k, n_table, n_database, paswrd,Entry_code_1, Entry_name_1,
Entry_code_2, Entry_name_2, identities, similarities):
```



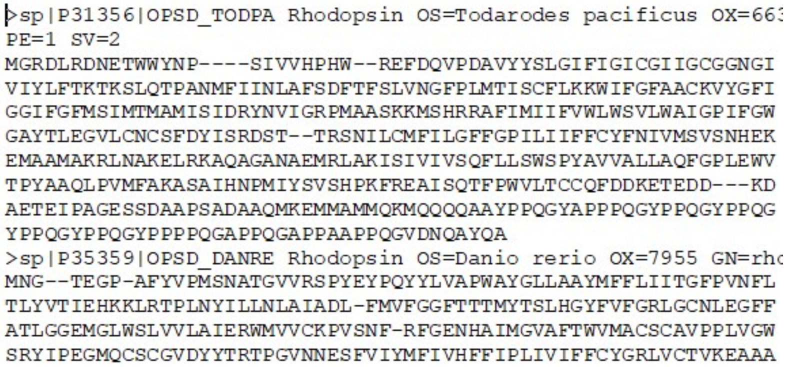
```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| aln                |
| mysql              |
| performance_schema |
| players            |
| sys                |
+--------------------+
```



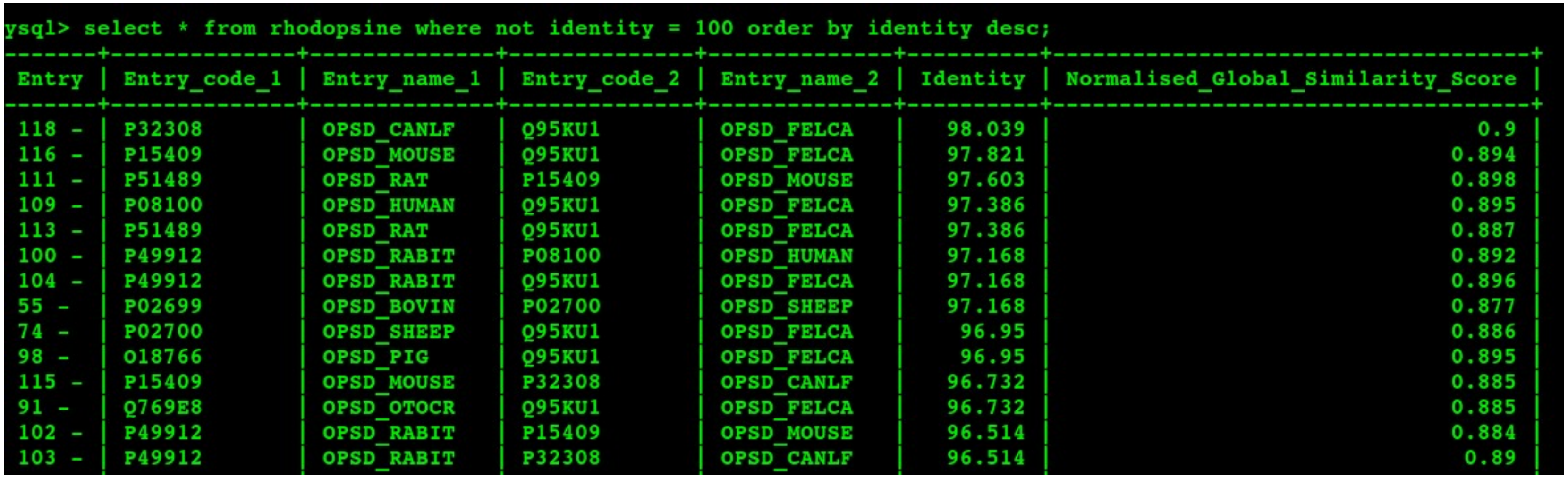
```
mysql> show tables;
+---------------+
| Tables_in_aln |
+---------------+
| rhodopsine    |
+---------------+
```

ISC OUTPUT

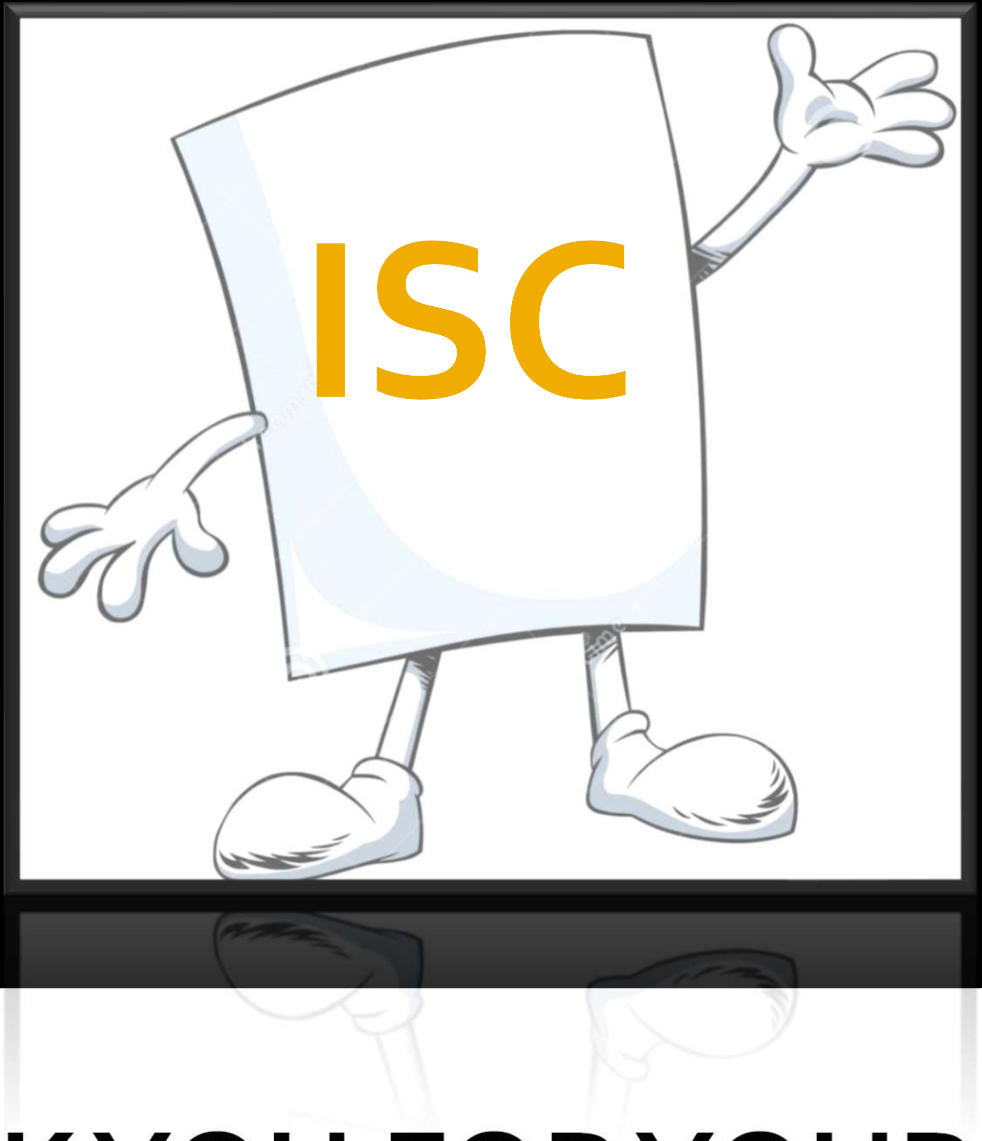
msa.fasta



```
>sp|P31356|OPSD_TODPA Rhodopsin OS=Todarodes pacificus OX=663
PE=1 SV=2
MGRDLRDNETWWYNP----SIVVHPHW--REFDQVPDAVYYSLGIFIGICGIIGCGGNGI
VIYLFTKTKSLQTPANMFIINLAFSDFTFSLVNGFPLMTISCFLKKWIFGFAACKVYGFI
GGIFGFMSIMTMAMISIDRYNVIGRPMAASKKMSHRRAFIMIIFVWLWSVLWAIGPIFGW
GAYTLEGVLCNCSFDYISRDST--TRSNILCMFILGFFGPILIIFFCYFNIVMSVSNHEK
EMAAMAKRLNAKELRKAQAGANAEMRLAKISIVIVSQFLLSWSPYAVVALLAQFGPLEWV
TPYAAQLPVMFAKASAIHNPMIYSVSHPKFREAISQTFPWVLTCCQFDDKETEDD---KD
AETEIPAGESSDAAPSADAAQMKEMMAMMQKMQQQQAAYPPQGYAPPPQGYPPQGYPPQG
YPPQGYPPQGYPPPPQGAPPQGAPPAAPPQGVDNQAYQA
>sp|P35359|OPSD_DANRE Rhodopsin OS=Danio rerio OX=7955 GN=rh
MNG--TEGP-AFYVPMSNATGVVRSPYEYPQYYLVAPWAYGLLAAYMFFLIITGFPVNFL
TLYVTIEHKKLRTPLNYILLNLAIADL-FMVFGGFTTTMYTSLHGYFVFGRLGCNLEGFF
ATLGGEMGLWSLVVLAIERWMVVCKPVSNF-RFGENHAIMGVAFTWVMACSCAVPPLVGW
SRYIPEGMQCSCGVDYYTRTPGVNNESFVIYMFIVHFFIPLIVIFFCYGRLVCTVKEAAA
```



```
ysql> select * from rhodopsine where not identity = 100 order by identity desc;
-------+-------------+-------------+-------------+-------------+-----------+---------------------------------+
 Entry | Entry_code_1 | Entry_name_1 | Entry_code_2 | Entry_name_2 | Identity | Normalised_Global_Similarity_Score |
-------+-------------+-------------+-------------+-------------+-----------+---------------------------------+
 118 - | P32308       | OPSD_CANLF   | Q95KU1       | OPSD_FELCA   |   98.039 |                                0.9 |
 116 - | P15409       | OPSD_MOUSE   | Q95KU1       | OPSD_FELCA   |   97.821 |                              0.894 |
 111 - | P51489       | OPSD_RAT     | P15409       | OPSD_MOUSE   |   97.603 |                              0.898 |
 109 - | P08100       | OPSD_HUMAN   | Q95KU1       | OPSD_FELCA   |   97.386 |                              0.895 |
 113 - | P51489       | OPSD_RAT     | Q95KU1       | OPSD_FELCA   |   97.386 |                              0.887 |
 100 - | P49912       | OPSD_RABIT   | P08100       | OPSD_HUMAN   |   97.168 |                              0.892 |
 104 - | P49912       | OPSD_RABIT   | Q95KU1       | OPSD_FELCA   |   97.168 |                              0.896 |
 55 -  | P02699       | OPSD_BOVIN   | P02700       | OPSD_SHEEP   |   97.168 |                              0.877 |
 74 -  | P02700       | OPSD_SHEEP   | Q95KU1       | OPSD_FELCA   |    96.95 |                              0.886 |
 98 -  | O18766       | OPSD_PIG     | Q95KU1       | OPSD_FELCA   |    96.95 |                              0.895 |
 115 - | P15409       | OPSD_MOUSE   | P32308       | OPSD_CANLF   |   96.732 |                              0.885 |
 91 -  | Q769E8       | OPSD_OTOCR   | Q95KU1       | OPSD_FELCA   |   96.732 |                              0.885 |
 102 - | P49912       | OPSD_RABIT   | P15409       | OPSD_MOUSE   |   96.514 |                              0.884 |
 103 - | P49912       | OPSD_RABIT   | P32308       | OPSD_CANLF   |   96.514 |                               0.89 |
```



THANK YOU FOR YOUR TIME!