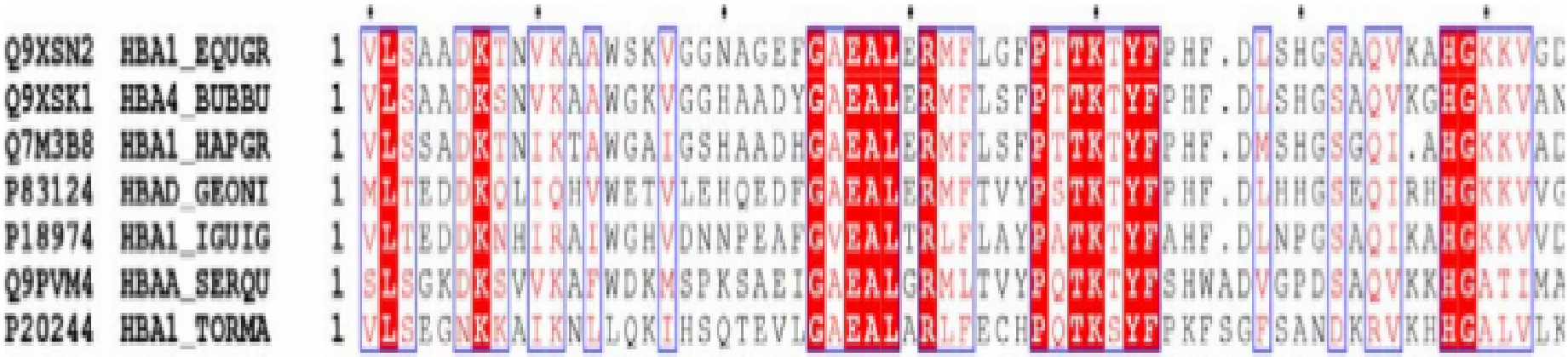
How do MSA algorithms compare?

CENG4525 Introduction to Bioinformatics

Aslıhan Arslan
Uğur Özcan
İsmail Uztemur

January 9, 2017

What is MSA?


Q9XSN2 HBA1_EQUGR 1
Q9XSK1 HBA4_BUBBU 1
Q7M3B8 HBA1_HAPGR 1
P83124 HBAD_GEONI 1
P18974 HBA1_IGUIG 1
Q9PVM4 HBAA_SERQU 1
P20244 HBA1_TORMA 1


RLA0_METVA
RLA0_METJA
RLA0_PYRAB
RLA0_PYRHO
RLA0_PYRFU
RLA0_PYRKO
RLA0_HALMA
RLA0_HALVO
RLA0_HALSA
RLA0_THEAC
RLA0_THEVO
RLA0_PICTO

About The Project

Our Research Question

- How do MSA algorithms compare?

Team Members and Their Roles

- Aslıhan Arslan - Testing Muscle and Graphical Analysis
- Uğur Özcan - Testing ClustalW and Creating Scripts
- Ismail Uztemur - Testing TCoffee and Writing Project Report

About The Project

Database/Benchmark

- Balibase 3.0

MSA Tools

- Muscle
- ClustalW
- TCoffee

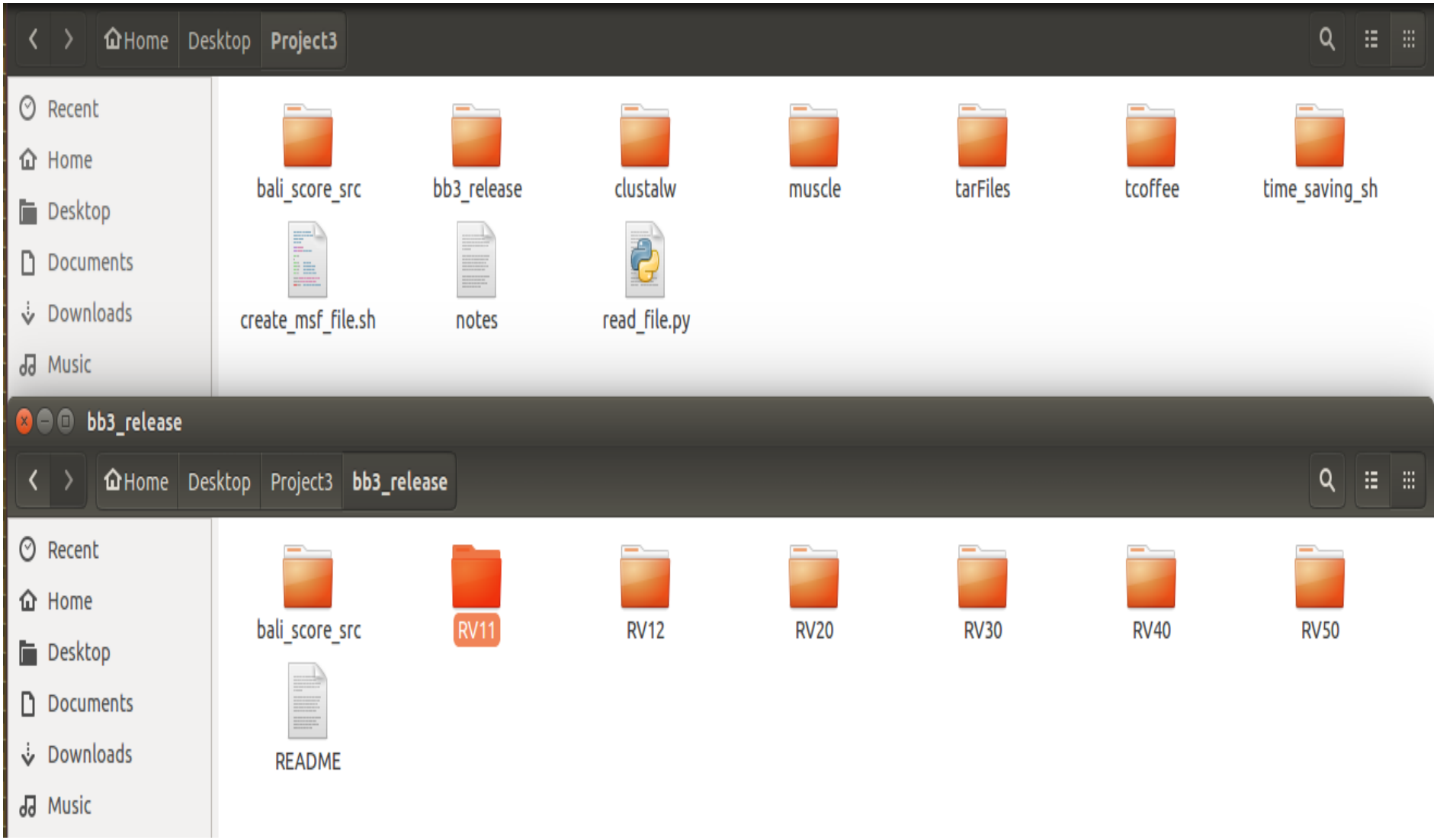
Programming Tool

- Python 3.5

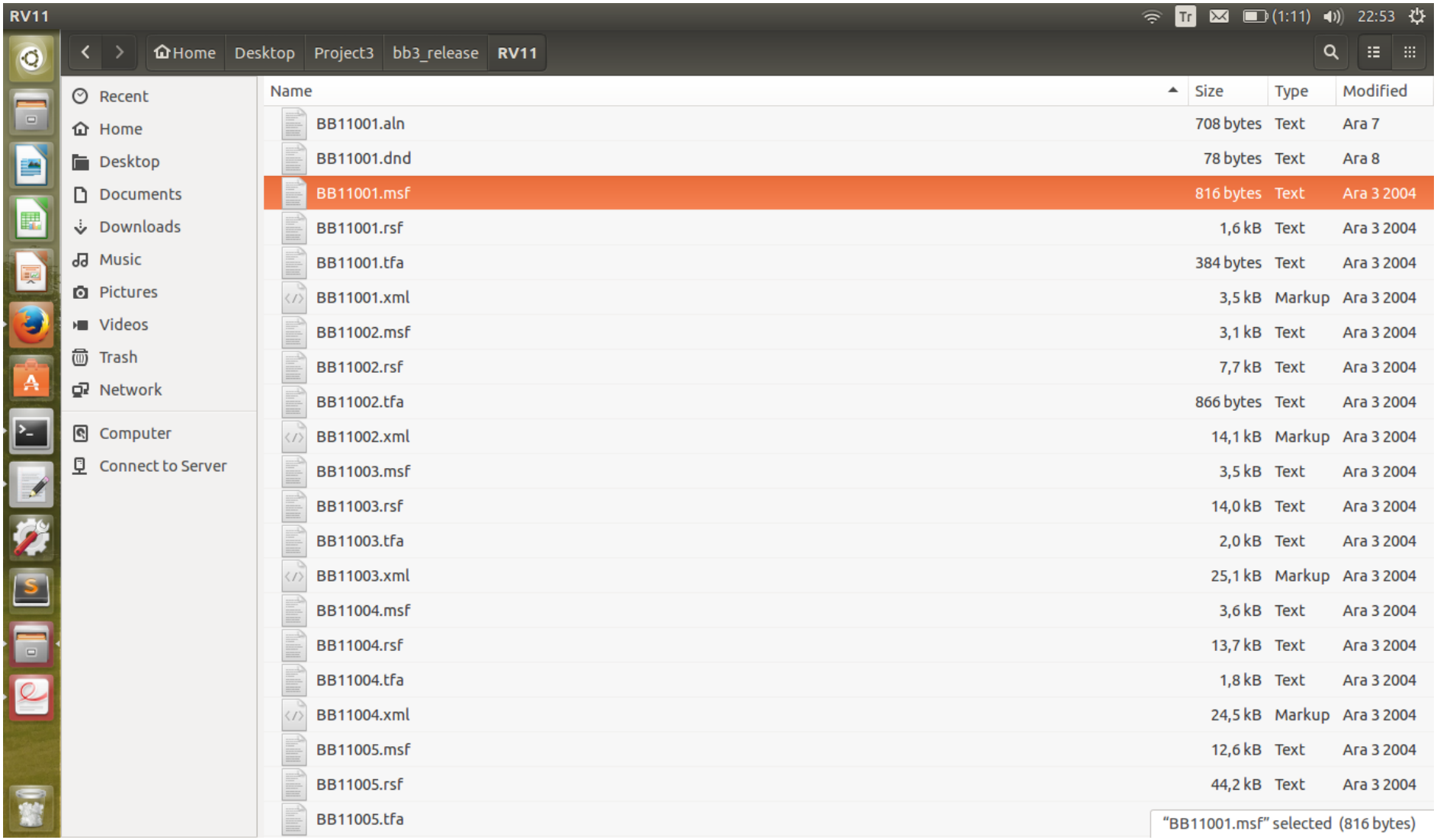
Project Files

```
*****************************************************************************

               BAliBASE multiple alignment benchmark suite
                        (version 3.0, Dec 2004)

*****************************************************************************


Please send bug reports, comments etc. to :-
        thompson@igbmc.u-strasbg.fr

*****************************************************************************

The tests in the BAliBASE benchmark suite are divided into 5 different reference sets.

Reference 1 equi-distant sequences with 2 different levels of conservation,

Reference 2 families aligned with a highly divergent "orphan" sequence,

Reference 3 subgroups with <25% residue identity between groups,

Reference 4 sequences with N/C-terminal extensions,

Reference 5 internal insertions.

The different reference sets are organised in seperate directories:

RV11:   Reference 1, very divergent sequences (<20% identity)
RV12:   Reference 1, medium to divergent sequences (20-40% identity)
RV20:   Reference 2
RV30:   Reference 3
RV40:   Reference 4
RV50:   Reference 5

For each test, a number of files are provided. The files named BBnnnnn contain the
full-length sequences, while the files named BBSnnnnn contain the sequences corresponding
to the homologous regions only. Different file formats are also provided:

BBnnnnn.xml     the XML file with the aligned full length sequences and the annotation
BBnnnnn.msf     the MSF file with the aligned full length sequences
BBnnnnn.tfa     the TFA file with the non-aligned full length sequences
BBSnnnnn.xml    the XML file with the aligned truncated sequences and the annotation
BBSnnnnn.msf    the MSF file with the aligned truncated sequences
BBSnnnnn.tfa    the TFA file with the non-aligned truncated sequences
```

Project Files



Project Files



Project Steps

```
steps to run project files
==========================

1- Download the "Project3.zip" to your PC.
2- Extract the zip file to Desktop and see "Project3" folder.
3- Open the Terminal and go to Project3 folder. (like "cd Desktop/Project3/")
4- Run the "create_msf_file.sh" file. (sh create_msf_file.sh). This sh file creates msf files in clustalw, muscle and tcoffee folder.
5- Go to bali_score_src folder. (cd bali_score_src/)
6- Run the "testBali.sh" file. (sh testBali.sh). This sh file creates bali_score file in clustalw, muscle and tcoffee folder. This script run
"bali_score" file at the background.
7- Get the results on Terminal

sample terminal codes
=====================

t_coffee /home/aslihan/Documents/Project3/bb3_release/RV11/BB11005.tfa -output=clustalw,msf

clustalw -infile=/home/aslihan/Documents/Project3/bb3_release/RV11/BB11005.tfa -output=gcg

muscle -in /home/aslihan/Documents/Project3/bb3_release/RV11/BB11005.tfa -out /home/aslihan/Documents/Project3/muscle/RV11/BB11005.msf -msf


./bali_score /home/aslihan/Desktop/Project3/bb3_release/RV11/BB11001.msf /home/aslihan/Desktop/Project3/tcoffee/RV11/BB11001.msf
'
```

Project Results

```
ugur@ugur:~/Desktop/Project3/bali_score_src$ sh testBali.sh
ugur
['RV11', 'BB11001.msf', 'tcoffee', '0.862', '0.770', 'clustalw', '0.961', '0.920', 'muscle', '0.904', '0.850']
['RV11', 'BB11002.msf', 'tcoffee', '0.537', '0.000', 'clustalw', '0.435', '0.000', 'muscle', '0.572', '0.270']
['RV11', 'BB11003.msf', 'tcoffee', '0.612', '0.430', 'clustalw', '0.609', '0.420', 'muscle', '0.561', '0.390']
['RV11', 'BB11004.msf', 'tcoffee', '0.625', '0.450', 'clustalw', '0.196', '0.000', 'muscle', '0.600', '0.420']
['RV11', 'BB11005.msf', 'tcoffee', '0.491', '0.140', 'clustalw', '0.377', '0.110', 'muscle', '0.381', '0.010']
['RV11', 'BB11006.msf', 'tcoffee', '0.409', '0.180', 'clustalw', '0.250', '0.000', 'muscle', '0.307', '0.070']
['RV11', 'BB11007.msf', 'tcoffee', '0.677', '0.420', 'clustalw', '0.557', '0.260', 'muscle', '0.593', '0.390']
['RV11', 'BB11008.msf', 'tcoffee', '0.541', '0.430', 'clustalw', '0.478', '0.410', 'muscle', '0.393', '0.230']
['RV11', 'BB11009.msf', 'tcoffee', '0.265', '0.000', 'clustalw', '0.295', '0.000', 'muscle', '0.316', '0.000']
['RV11', 'BB11010.msf', 'tcoffee', '0.354', '0.170', 'clustalw', '0.201', '0.000', 'muscle', '0.245', '0.060']
['RV11', 'BB11011.msf', 'tcoffee', '0.264', '0.160', 'clustalw', '0.198', '0.070', 'muscle', '0.145', '0.000']
['RV11', 'BB11012.msf', 'tcoffee', '0.925', '0.880', 'clustalw', '0.910', '0.850', 'muscle', '0.869', '0.770']
['RV11', 'BB11013.msf', 'tcoffee', '0.143', '0.000', 'clustalw', '0.084', '0.000', 'muscle', '0.076', '0.000']
['RV11', 'BB11014.msf', 'tcoffee', '0.798', '0.660', 'clustalw', '0.707', '0.530', 'muscle', '0.756', '0.600']
['RV11', 'BB11015.msf', 'tcoffee', '0.677', '0.480', 'clustalw', '0.654', '0.430', 'muscle', '0.828', '0.750']
['RV11', 'BB11016.msf', 'tcoffee', '0.496', '0.090', 'clustalw', '0.367', '0.000', 'muscle', '0.407', '0.000']
['RV11', 'BB11017.msf', 'tcoffee', '0.790', '0.700', 'clustalw', '0.653', '0.480', 'muscle', '0.732', '0.610']
['RV11', 'BB11018.msf', 'tcoffee', '0.674', '0.330', 'clustalw', '0.440', '0.150', 'muscle', '0.532', '0.230']
['RV11', 'BB11019.msf', 'tcoffee', '0.687', '0.210', 'clustalw', '0.505', '0.090', 'muscle', '0.646', '0.220']
['RV11', 'BB11020.msf', 'tcoffee', '0.693', '0.320', 'clustalw', '0.652', '0.320', 'muscle', '0.592', '0.300']
['RV11', 'BB11021.msf', 'tcoffee', '0.694', '0.530', 'clustalw', '0.173', '0.000', 'muscle', '0.562', '0.450']
['RV11', 'BB11022.msf', 'tcoffee', '0.082', '0.000', 'clustalw', '0.251', '0.000', 'muscle', '0.073', '0.000']
['RV11', 'BB11023.msf', 'tcoffee', '0.508', '0.250', 'clustalw', '0.373', '0.150', 'muscle', '0.420', '0.130']
['RV11', 'BB11024.msf', 'tcoffee', '0.474', '0.220', 'clustalw', '0.173', '0.000', 'muscle', '0.197', '0.000']
['RV11', 'BB11025.msf', 'tcoffee', '0.173', '0.000', 'clustalw', '0.118', '0.000', 'muscle', '0.092', '0.000']
['RV11', 'BB11026.msf', 'tcoffee', '0.225', '0.000', 'clustalw', '0.250', '0.000', 'muscle', '0.305', '0.000']
['RV11', 'BB11027.msf', 'tcoffee', '0.433', '0.150', 'clustalw', '0.253', '0.000', 'muscle', '0.380', '0.170']
['RV11', 'BB11028.msf', 'tcoffee', '0.563', '0.000', 'clustalw', '0.476', '0.000', 'muscle', '0.402', '0.000']
['RV11', 'BB11029.msf', 'tcoffee', '0.505', '0.430', 'clustalw', '0.505', '0.460', 'muscle', '0.432', '0.330']
['RV11', 'BB11030.msf', 'tcoffee', '0.629', '0.210', 'clustalw', '0.337', '0.110', 'muscle', '0.452', '0.100']
['RV11', 'BB11031.msf', 'tcoffee', '0.582', '0.200', 'clustalw', '0.287', '0.000', 'muscle', '0.366', '0.000']
['RV11', 'BB11032.msf', 'tcoffee', '0.728', '0.410', 'clustalw', '0.520', '0.160', 'muscle', '0.630', '0.300']
['RV11', 'BB11033.msf', 'tcoffee', '0.467', '0.000', 'clustalw', '0.456', '0.000', 'muscle', '0.367', '0.000']
['RV11', 'BB11034.msf', 'tcoffee', '0.512', '0.110', 'clustalw', '0.277', '0.000', 'muscle', '0.419', '0.000']
['RV11', 'BB11035.msf', 'tcoffee', '0.508', '0.340', 'clustalw', '0.468', '0.300', 'muscle', '0.592', '0.430']
['RV11', 'BB11036.msf', 'tcoffee', '0.575', '0.240', 'clustalw', '0.480', '0.130', 'muscle', '0.538', '0.130']
['RV11', 'BB11037.msf', 'tcoffee', '0.513', '0.300', 'clustalw', '0.364', '0.110', 'muscle', '0.333', '0.120']
['RV11', 'BB11038.msf', 'tcoffee', '0.809', '0.610', 'clustalw', '0.485', '0.000', 'muscle', '0.664', '0.360']
['RV11', 'BBS11001.msf', 'tcoffee', '0.934', '0.890', 'clustalw', '0.974', '0.940', 'muscle', '0.993', '0.980']
['RV11', 'BBS11002.msf', 'tcoffee', '0.828', '0.480', 'clustalw', '0.731', '0.420', 'muscle', '0.715', '0.480']
['RV11', 'BBS11003.msf', 'tcoffee', '0.723', '0.540', 'clustalw', '0.605', '0.380', 'muscle', '0.695', '0.510']
['RV11', 'BBS11004.msf', 'tcoffee', '0.645', '0.480', 'clustalw', '0.449', '0.200', 'muscle', '0.561', '0.410']
['RV11', 'BBS11005.msf', 'tcoffee', '0.518', '0.170', 'clustalw', '0.401', '0.120', 'muscle', '0.476', '0.160']
```

SP and TC Scores

Two different scores estimate the accuracy of an alignment comparing to the BAliBASE reference alignment: the Sum-of-Pairs score (SP), and the Total-Column score (TC)[1].

Example 1.

Reference alignment		Tested alignment	
s01	ABCDE	s01	ABCDE.
s02	A...B	s02	A....B
s03	AB..C	s03	AB..C.
s04	ABC.D	s04	ABC.D.
s05	ABCDE	s05	ABCDE.
s06	ABCDE	s06	ABCDE.
s07	ABCDE	s07	ABCDE.
s08	ABCDE	s08	ABCDE.
s09	ABCDE	s09	ABCDE.
s10	....A	s10	....A.

SP. There are the following pairs in columns in the reference alignment:
1st column: $9 \cdot 8/2 = 36$,
2nd column: $8 \cdot 7/2 = 28$,
3rd column: $7 \cdot 6/2 = 21$,
4th column: $6 \cdot 5/2 = 15$,
5th column: $10 \cdot 9/2 = 45$,
which results in the total number of pairs in the reference alignment, equal to 145. In the tested alignment the first four columns are the same as in the reference alignment and in the fifth column there are $9 \cdot 8/2 = 36$ correct pairs. So, the total number of correct pairs in the tested alignment is equal to 136, and thus $SP = 136/145 = 0.938$.

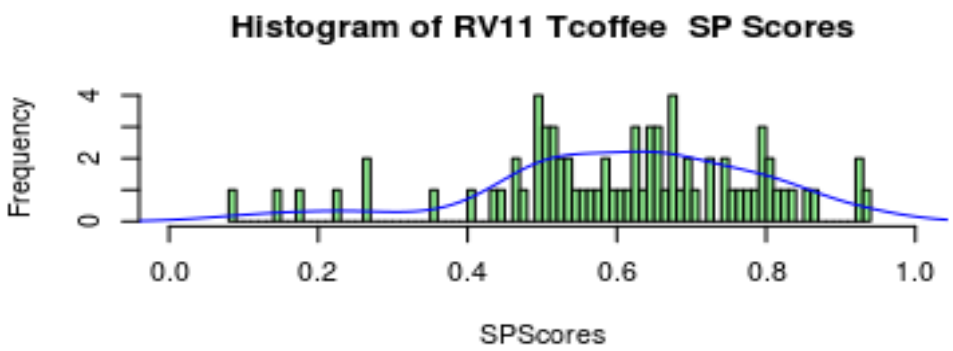
TC. Since four columns are correct, $TC = 4/5 = 0.8$.

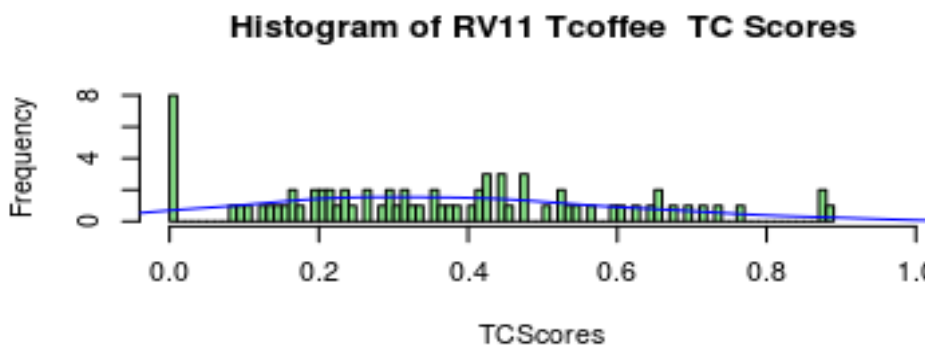
Figure:SP TC Scores [2].

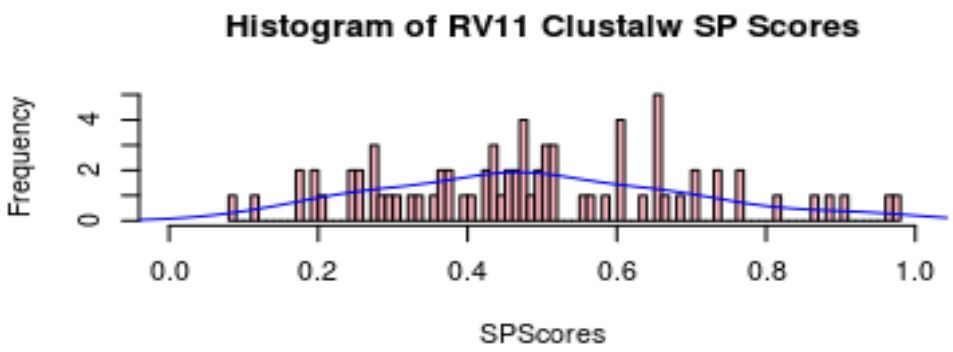
Summary of RV11 file Scores in R

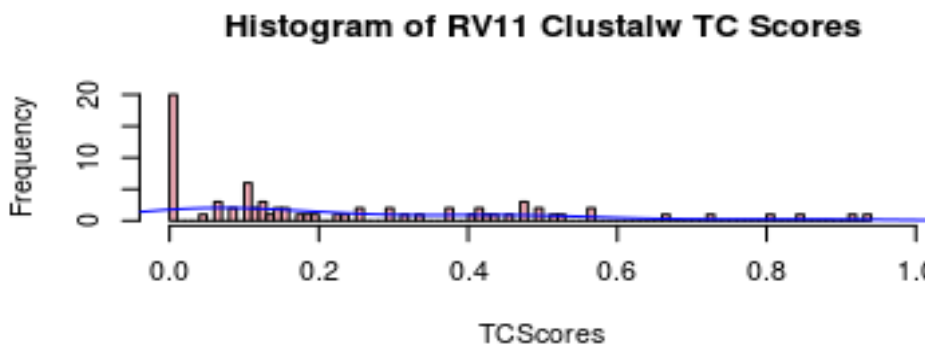
```
Console ~/
> summary(RV11)
  Folder             File         tcoffee     tcoffeeSP         tcoffeeTC           clustalw
 RV11:76    BB11001.msf: 1    tcoffee:76   Min.   :0.0820   Min.   :0.0000    clustalw:76
            BB11002.msf: 1                 1st Qu.:0.5080   1st Qu.:0.2000
            BB11003.msf: 1                 Median :0.6270   Median :0.3500
            BB11004.msf: 1                 Mean   :0.6071   Mean   :0.3634
            BB11005.msf: 1                 3rd Qu.:0.7320   3rd Qu.:0.5150
            BB11006.msf: 1                 Max.   :0.9340   Max.   :0.8900
           (Other)     :70
   clustalwSP       clustalwTC          muscle      muscleSP         muscleTC
 Min.   :0.0840   Min.   :0.0000    muscle:76   Min.   :0.0730   Min.   :0.0000
 1st Qu.:0.3528   1st Qu.:0.0000                1st Qu.:0.4198   1st Qu.:0.1150
 Median :0.4770   Median :0.1450                Median :0.5585   Median :0.2700
 Mean   :0.4908   Mean   :0.2395                Mean   :0.5479   Mean   :0.3125
 3rd Qu.:0.6415   3rd Qu.:0.4200                3rd Qu.:0.6883   3rd Qu.:0.4525
 Max.   :0.9740   Max.   :0.9400                Max.   :0.9930   Max.   :0.9800
```

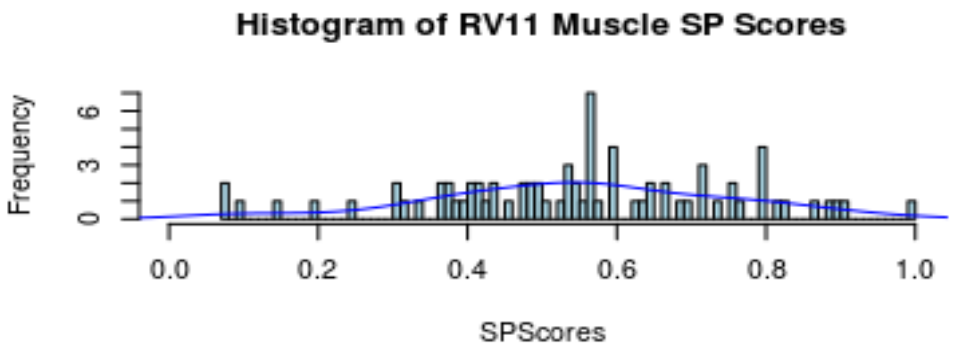
Histograms of RV11

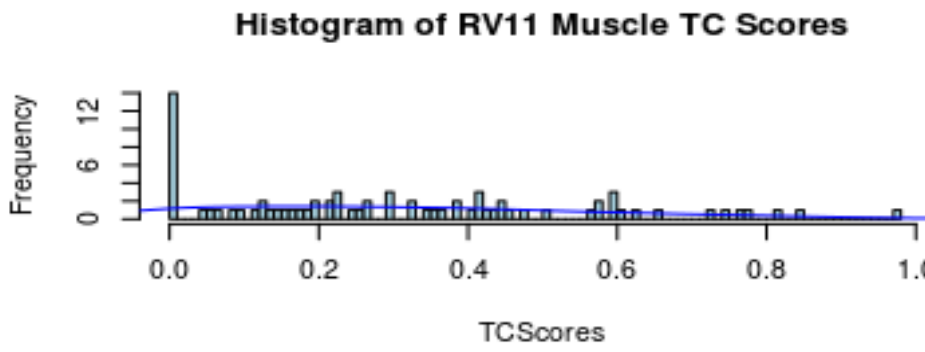










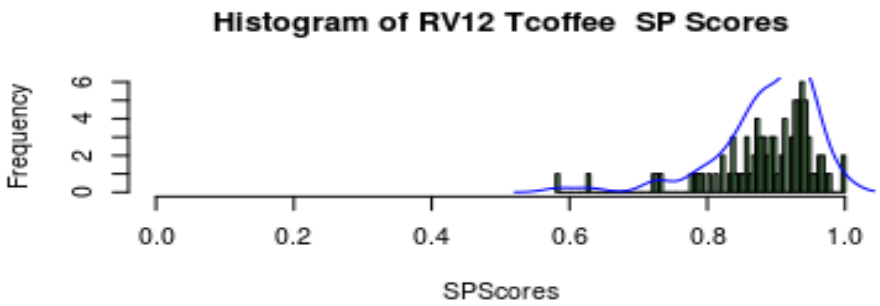


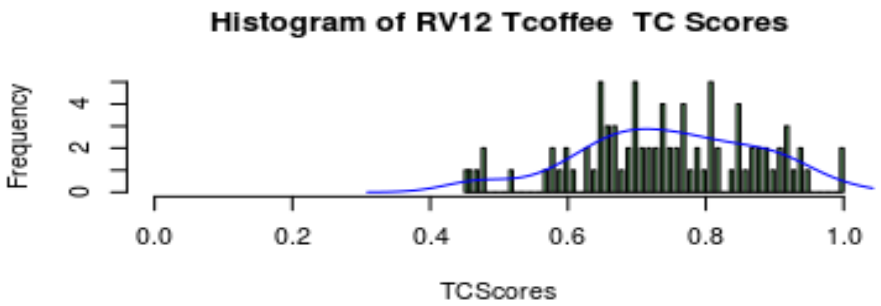
Summary of RV12 file Scores in R

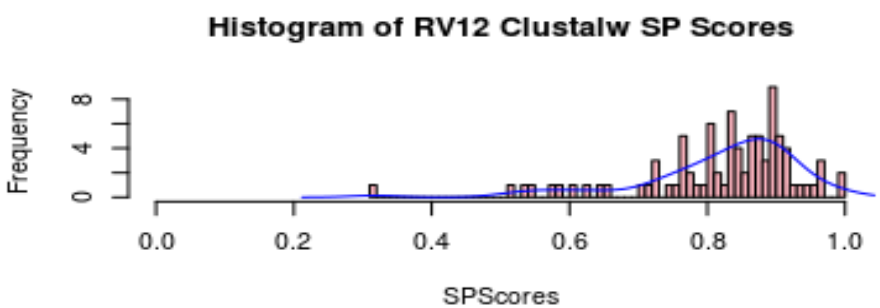
Console ~/

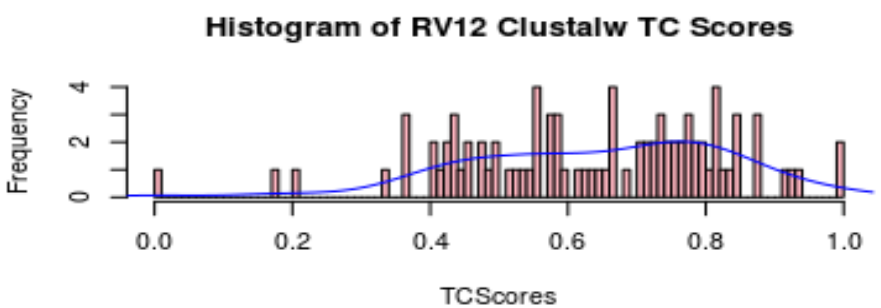
```
> summary(RV12)
  Folder              File         tcoffee      tcoffeeSP         tcoffeeTC           clustalw
 RV12:88    BB12001.msf: 1    tcoffee:88    Min.   :0.5840    Min.   :0.4500    clustalw:88
            BB12002.msf: 1                  1st Qu.:0.8598    1st Qu.:0.6600
            BB12003.msf: 1                  Median :0.9000    Median :0.7400
            BB12004.msf: 1                  Mean   :0.8881    Mean   :0.7448
            BB12005.msf: 1                  3rd Qu.:0.9370    3rd Qu.:0.8500
            BB12006.msf: 1                  Max.   :1.0000    Max.   :1.0000
           (Other)     :82
   clustalwSP       clustalwTC         muscle      muscleSP         muscleTC
 Min.   :0.3170   Min.   :0.0000    muscle:88   Min.   :0.5500   Min.   :0.3900
 1st Qu.:0.7705   1st Qu.:0.5000                1st Qu.:0.8245   1st Qu.:0.6100
 Median :0.8415   Median :0.6700                Median :0.8830   Median :0.7100
 Mean   :0.8193   Mean   :0.6420                Mean   :0.8578   Mean   :0.6977
 3rd Qu.:0.8980   3rd Qu.:0.7825                3rd Qu.:0.9200   3rd Qu.:0.8000
 Max.   :1.0000   Max.   :1.0000                Max.   :1.0000   Max.   :1.0000
```

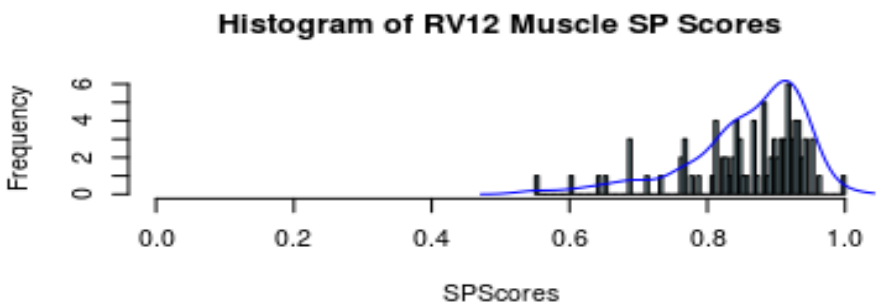
Histograms of RV12

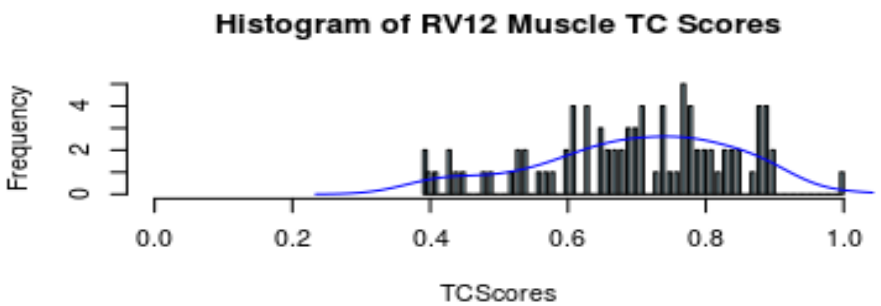








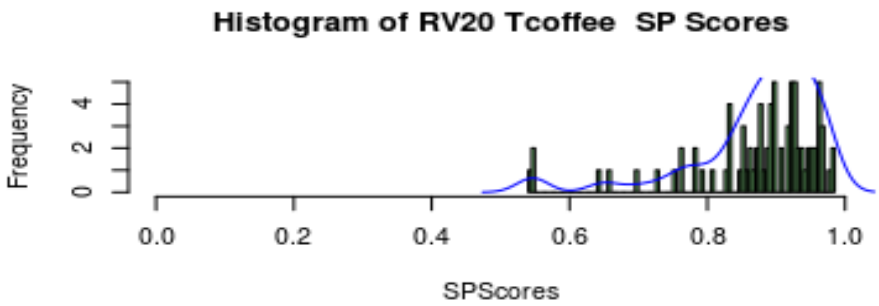


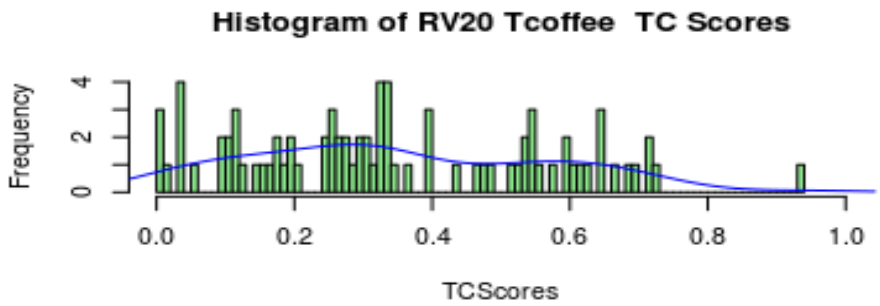


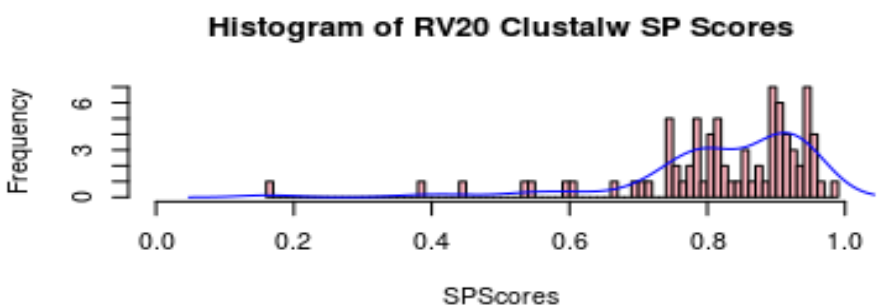
Summary of RV20 file Scores in R

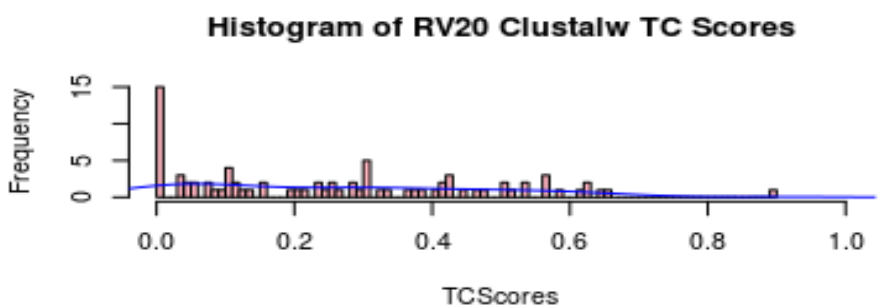
```
Console ~/
> summary(RV20)
  Folder              File         tcoffee      tcoffeeSP         tcoffeeTC           clustalw
 RV20:82    BB20001.msf: 1    tcoffee:82    Min.   :0.5420    Min.   :0.0000    clustalw:82
            BB20002.msf: 1                  1st Qu.:0.8512    1st Qu.:0.1800
            BB20003.msf: 1                  Median :0.8960    Median :0.3250
            BB20004.msf: 1                  Mean   :0.8744    Mean   :0.3485
            BB20005.msf: 1                  3rd Qu.:0.9327    3rd Qu.:0.5400
            BB20006.msf: 1                  Max.   :0.9830    Max.   :0.9400
           (Other)     :76
   clustalwSP       clustalwTC          muscle       muscleSP          muscleTC
 Min.   :0.1640   Min.   :0.0000    muscle:82   Min.   :0.2170   Min.   :0.000
 1st Qu.:0.7795   1st Qu.:0.0600                1st Qu.:0.8213   1st Qu.:0.120
 Median :0.8525   Median :0.2450                Median :0.8810   Median :0.280
 Mean   :0.8219   Mean   :0.2602                Mean   :0.8527   Mean   :0.303
 3rd Qu.:0.9187   3rd Qu.:0.4275                3rd Qu.:0.9285   3rd Qu.:0.450
 Max.   :0.9870   Max.   :0.9000                Max.   :0.9760   Max.   :0.900
```

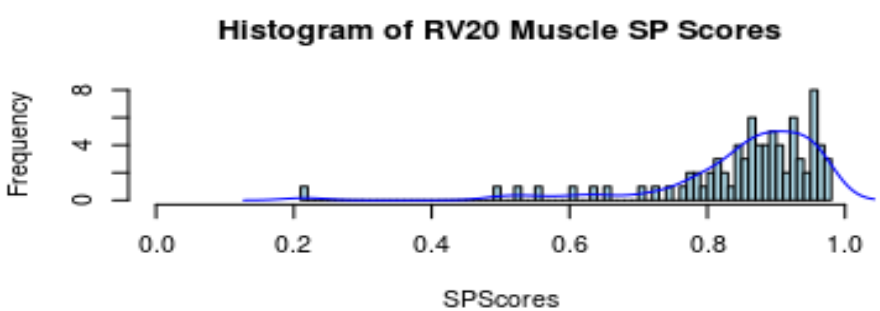
Histograms of RV20

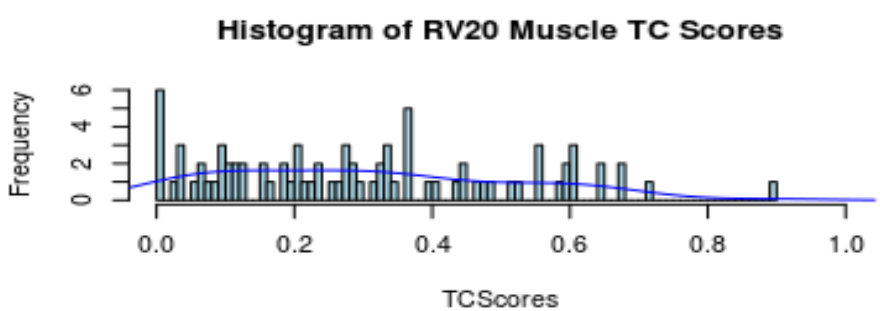








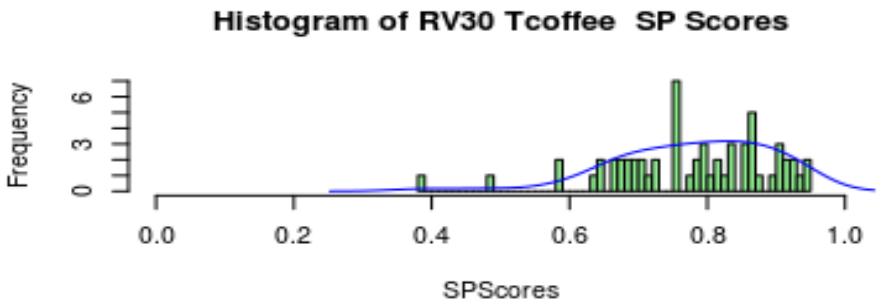


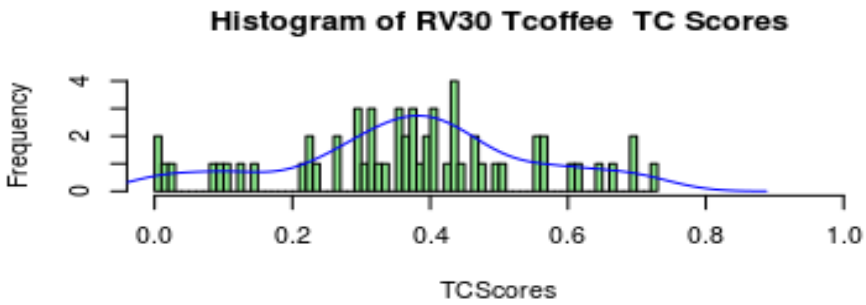


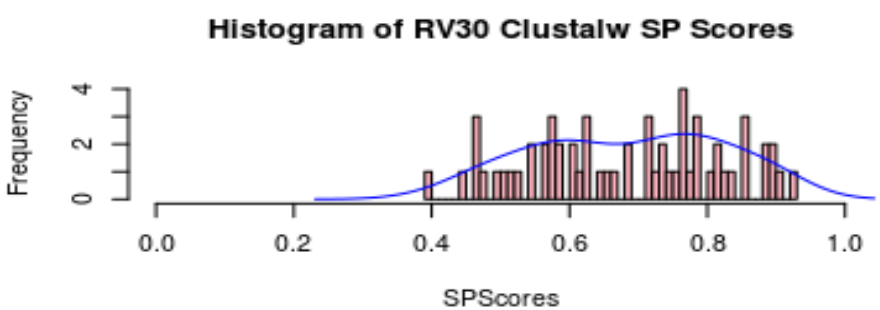
Summary of RV30 file Scores in R

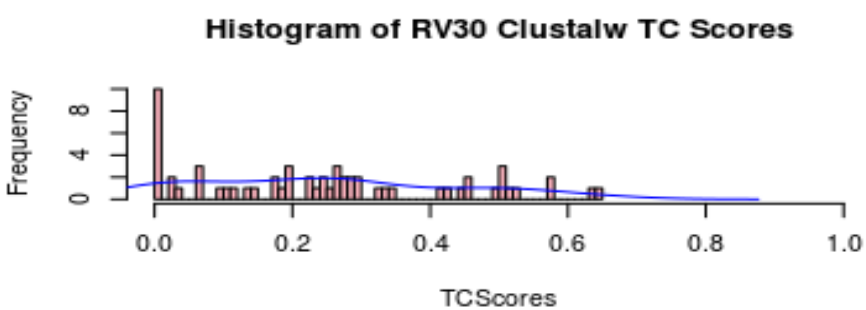
```
Console ~/
> summary(RV30)
  Folder              File         tcoffee      tcoffeeSP         tcoffeeTC           clustalw
 RV30:60    BB30001.msf: 1    tcoffee:60   Min.   :0.3890   Min.   :0.0000    clustalw:60
            BB30002.msf: 1                 1st Qu.:0.7005   1st Qu.:0.2925
            BB30003.msf: 1                 Median :0.7910   Median :0.3800
            BB30004.msf: 1                 Mean   :0.7785   Mean   :0.3720
            BB30005.msf: 1                 3rd Qu.:0.8652   3rd Qu.:0.4700
            BB30006.msf: 1                 Max.   :0.9490   Max.   :0.7300
           (Other)     :54
   clustalwSP       clustalwTC         muscle      muscleSP         muscleTC
 Min.   :0.3960   Min.   :0.0000    muscle:60   Min.   :0.3750   Min.   :0.0000
 1st Qu.:0.5783   1st Qu.:0.0700                1st Qu.:0.6470   1st Qu.:0.1775
 Median :0.6995   Median :0.2450                Median :0.7580   Median :0.3150
 Mean   :0.6842   Mean   :0.2478                Mean   :0.7505   Mean   :0.3090
 3rd Qu.:0.7867   3rd Qu.:0.3675                3rd Qu.:0.8413   3rd Qu.:0.4525
 Max.   :0.9250   Max.   :0.6500                Max.   :0.9360   Max.   :0.6900
```

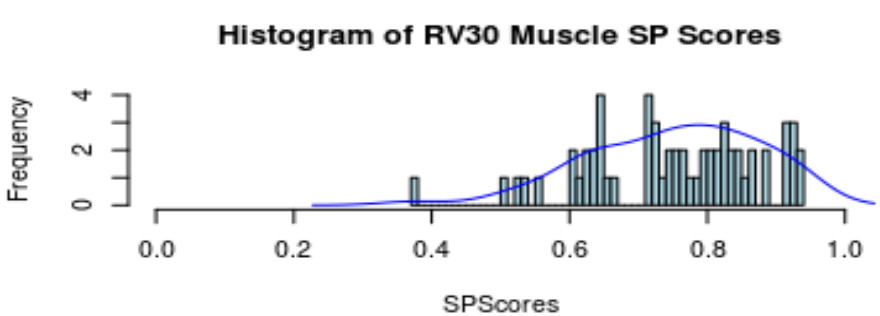
Histograms of RV30

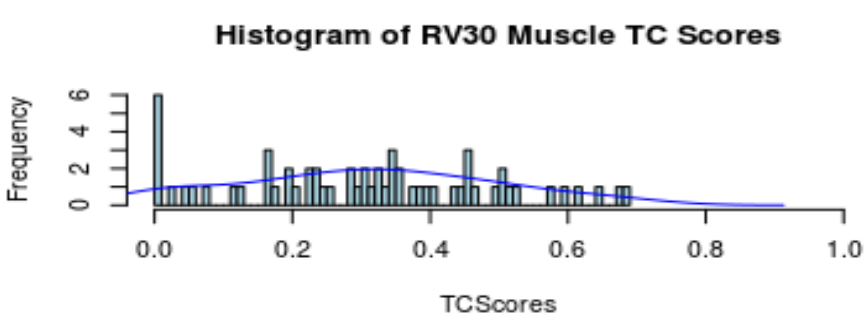










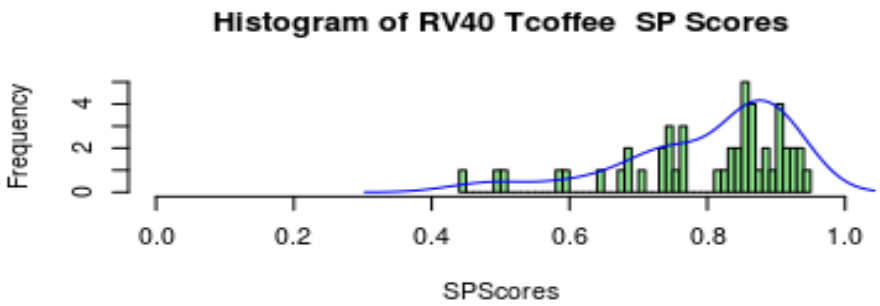


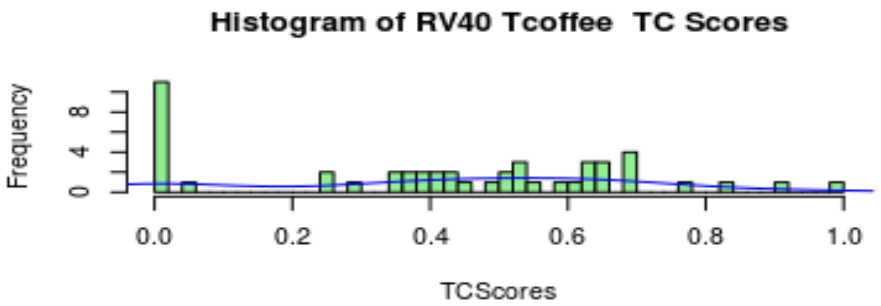
Summary of RV40 file Scores in R

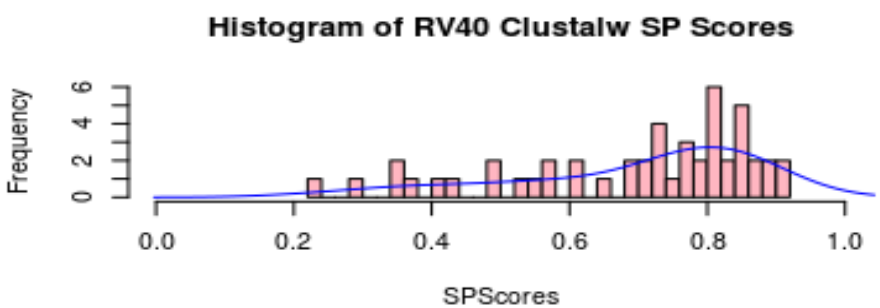
Console ~/

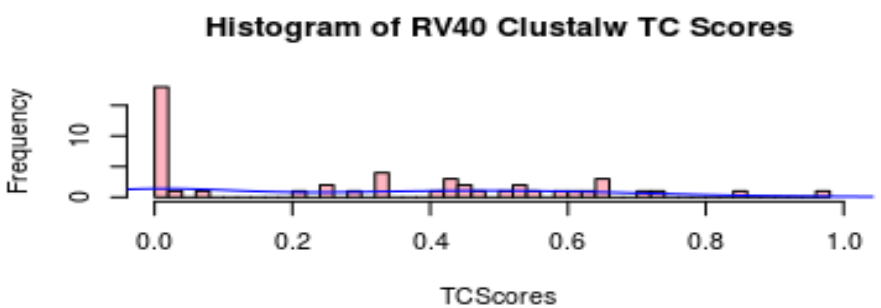
```
> summary(RV40)
  Folder             File         tcoffee     tcoffeeSP         tcoffeeTC           clustalw
 RV40:49    BB40001.msf: 1    tcoffee:49   Min.   :0.4400   Min.   :0.0000    clustalw:49
            BB40002.msf: 1                 1st Qu.:0.7410   1st Qu.:0.2600
            BB40003.msf: 1                 Median :0.8420   Median :0.4400
            BB40004.msf: 1                 Mean   :0.7996   Mean   :0.4155
            BB40005.msf: 1                 3rd Qu.:0.8890   3rd Qu.:0.6400
            BB40006.msf: 1                 Max.   :0.9440   Max.   :0.9900
           (Other)     :43
   clustalwSP       clustalwTC          muscle      muscleSP         muscleTC
 Min.   :0.2210   Min.   :0.0000    muscle:49   Min.   :0.4160   Min.   :0.0000
 1st Qu.:0.5780   1st Qu.:0.0000                1st Qu.:0.6740   1st Qu.:0.0000
 Median :0.7460   Median :0.3300                Median :0.8200   Median :0.3800
 Mean   :0.6951   Mean   :0.3024                Mean   :0.7593   Mean   :0.3384
 3rd Qu.:0.8240   3rd Qu.:0.5300                3rd Qu.:0.8860   3rd Qu.:0.5600
 Max.   :0.9200   Max.   :0.9800                Max.   :0.9330   Max.   :0.9600
```

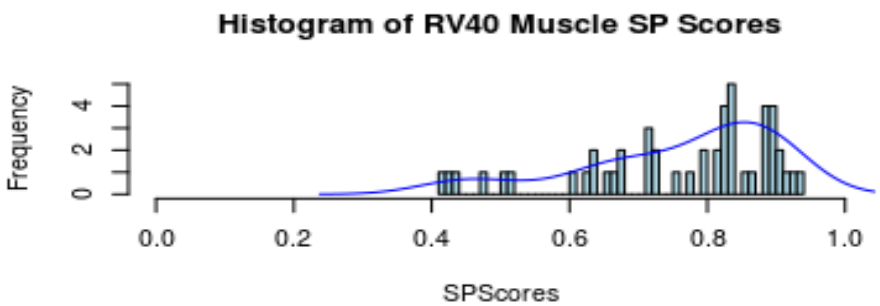
Histograms of RV40

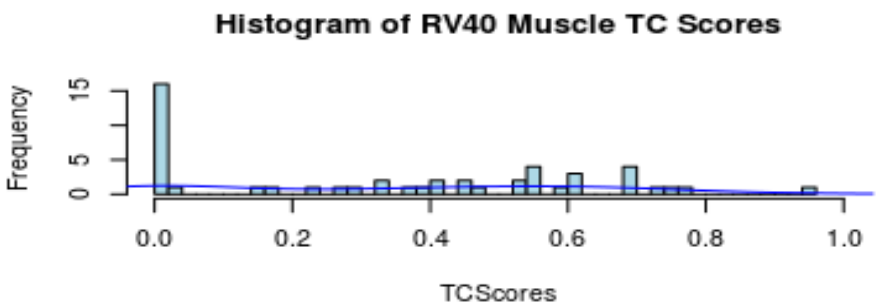








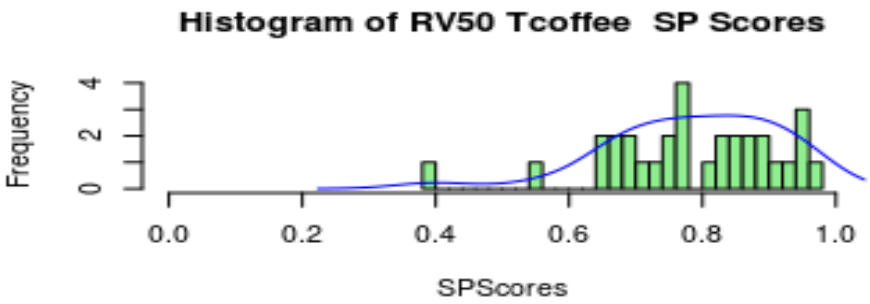


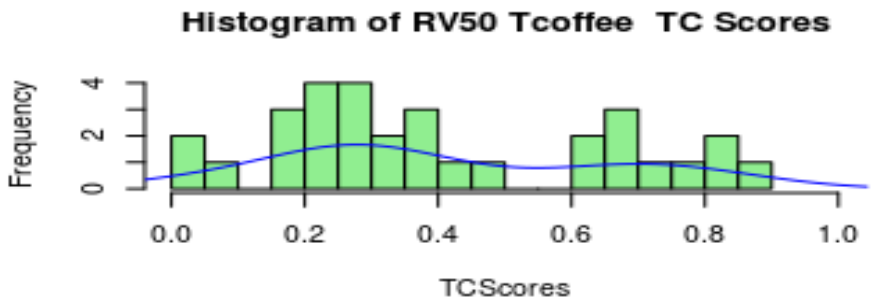


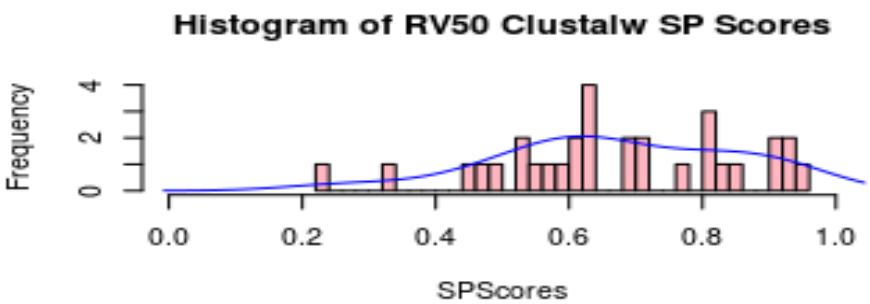
Summary of RV50 file Scores in R

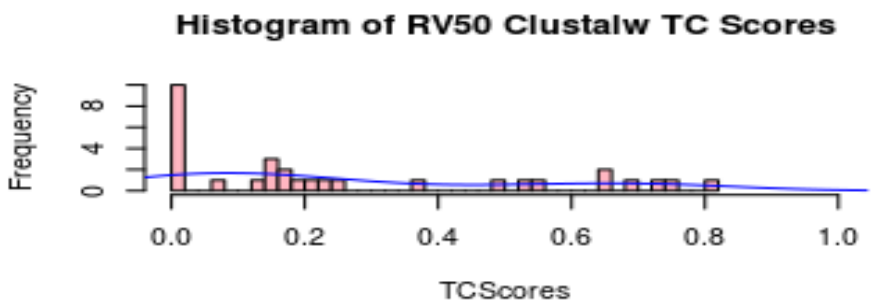
```
Console ~/
> summary(RV50)
  Folder             File        tcoffee     tcoffeeSP        tcoffeeTC          clustalw
 RV50:31    BB50001.msf: 1    tcoffee:31   Min.   :0.3980   Min.   :0.0000    clustalw:31
            BB50002.msf: 1                 1st Qu.:0.7045   1st Qu.:0.2500
            BB50003.msf: 1                 Median :0.7790   Median :0.3500
            BB50004.msf: 1                 Mean   :0.7862   Mean   :0.4116
            BB50005.msf: 1                 3rd Qu.:0.8805   3rd Qu.:0.6550
            BB50006.msf: 1                 Max.   :0.9620   Max.   :0.8800
           (Other)     :25
   clustalwSP       clustalwTC         muscle      muscleSP         muscleTC
 Min.   :0.2370   Min.   :0.0000    muscle:31   Min.   :0.2700   Min.   :0.0000
 1st Qu.:0.5580   1st Qu.:0.0050                1st Qu.:0.6265   1st Qu.:0.1000
 Median :0.6320   Median :0.1700                Median :0.7570   Median :0.3000
 Mean   :0.6699   Mean   :0.2658                Mean   :0.7318   Mean   :0.3245
 3rd Qu.:0.8115   3rd Qu.:0.5150                3rd Qu.:0.8380   3rd Qu.:0.5500
 Max.   :0.9450   Max.   :0.8100                Max.   :0.9490   Max.   :0.8400
```

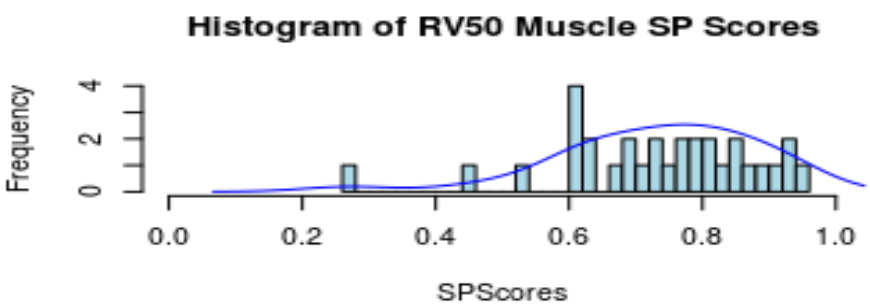
Histograms of RV50

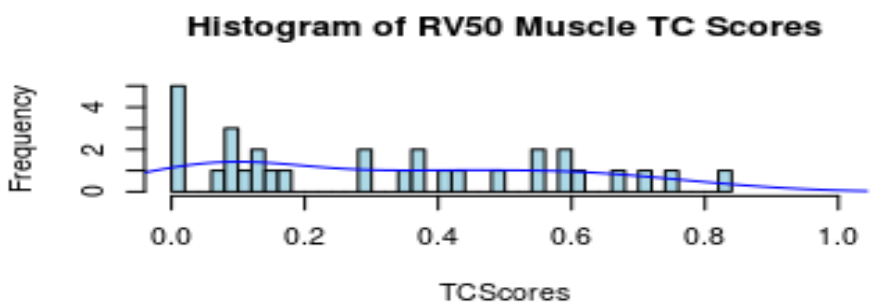












Results

- According to scores and histograms we can say that Tcoffee is more successful than Muscle.
- Muscle is more succesful than the Clustalw.

References

- 1.BŁAŻEWICZ J., FORMANOWICZ P.,WOJCIECHOWSKI P., SOME REMARKS ON EVALUATING THE QUALITY OF THE MULTIPLE SEQUENCE ALIGNMENT BASED ON THE BALIBASE BENCHMARK,2009
- 2. Pais FS, de Ruy P, Oliveira G, Coimbra R. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol. 2014;9(1):4. doi: 10.1186/1748-7188-9-4. [PMC free article] [PubMed] [Cross Ref]


thanks
for your
attention