Working with Alignments

In [1]:
```
# sudo apt-get update -y
# sudo apt-get install -y ugene

# sudo apt install prank
```

With UGENE installation also several tools are installed:

- clustalw
- clustalo
- muscle
- mafft
- t-coffee

In [2]:
```
import Bio.SeqIO as IO
```

Info about input data and programms

Input data

- **Ogranism:** Saccharomyces cerevisiae
- **Gene:** SUP35
- **Protein:** Eukaryotic peptide chain release factor GTP-binding subunit

Programs:

```
- ClustalW v2.1
- Clustal Omega v1.2.4
- MUSCLE v3.8.1551
- MAFFT v7.453
- Kalign v2.04
- t-coffee v11.0.8
- prank v.170427
- EMBOSS:6.6.0.0
```

Launch different alignment algorithms on the file SUP35_10seqs.fa

`SUP35_10seqs.fa` contains 10 sequences of the gene SUP35 from different S. cerevisiae strains

Note: it's better to save all files in fasta format for easier future comparison results from different programs with each other

ClustalW

Launch ClustalW with minimum parameters

In [3]:
```
! time clustalw -INFILE=SUP35_10seqs.fa
```

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: SUP35_Kla_AB039749                      2103 bp
Sequence 2: SUP35_Agos_ATCC_10895_NM_211584         2076 bp
Sequence 3: SUP35_Scer_74-D694_GCA_001578265.1      2058 bp
Sequence 4: SUP35_Spar_A12_Liti                     2046 bp
Sequence 5: SUP35_Smik_IFO1815T_30                  2046 bp
Sequence 6: SUP35_Skud_IFO1802T_36                  2037 bp
Sequence 7: SUP35_Sbou_unique28_CM003560            2058 bp
Sequence 8: SUP35_Scer_beer078_CM005938             2001 bp
Sequence 9: SUP35_Sarb_H-6_chrXIII_CM001575         2043 bp
Sequence 10: SUP35_Seub_CBS12357_chr_II_IV_DF968535  2025 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  70
Sequences (1:3) Aligned. Score:  71
Sequences (1:4) Aligned. Score:  71
Sequences (1:5) Aligned. Score:  71
Sequences (1:6) Aligned. Score:  71
Sequences (1:7) Aligned. Score:  71
Sequences (1:8) Aligned. Score:  69
Sequences (1:9) Aligned. Score:  62
Sequences (1:10) Aligned. Score:  73
Sequences (2:3) Aligned. Score:  66
Sequences (2:4) Aligned. Score:  69
Sequences (2:5) Aligned. Score:  69
Sequences (2:6) Aligned. Score:  69
Sequences (2:7) Aligned. Score:  66
Sequences (2:8) Aligned. Score:  70
Sequences (2:9) Aligned. Score:  69
Sequences (2:10) Aligned. Score:  69
Sequences (3:4) Aligned. Score:  93
Sequences (3:5) Aligned. Score:  88
Sequences (3:6) Aligned. Score:  85
Sequences (3:7) Aligned. Score:  99
Sequences (3:8) Aligned. Score:  99
Sequences (3:9) Aligned. Score:  86
Sequences (3:10) Aligned. Score:  84
Sequences (4:5) Aligned. Score:  89
Sequences (4:6) Aligned. Score:  86
Sequences (4:7) Aligned. Score:  93
Sequences (4:8) Aligned. Score:  93
Sequences (4:9) Aligned. Score:  87
Sequences (4:10) Aligned. Score:  84
Sequences (5:6) Aligned. Score:  85
Sequences (5:7) Aligned. Score:  88
Sequences (5:8) Aligned. Score:  88
Sequences (5:9) Aligned. Score:  86
Sequences (5:10) Aligned. Score:  84
Sequences (6:7) Aligned. Score:  85
Sequences (6:8) Aligned. Score:  85
Sequences (6:9) Aligned. Score:  86
Sequences (6:10) Aligned. Score:  85
Sequences (7:8) Aligned. Score:  99
Sequences (7:9) Aligned. Score:  86
Sequences (7:10) Aligned. Score:  84
Sequences (8:9) Aligned. Score:  85
Sequences (8:10) Aligned. Score:  84
Sequences (9:10) Aligned. Score:  86
Guide tree file created:   [SUP35_10seqs.dnd]

There are 9 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:37600
Group 2: Sequences:   3      Score:38275
Group 3: Sequences:   4      Score:36568
Group 4: Sequences:   5      Score:35737
Group 5: Sequences:   6      Score:35113
Group 6: Sequences:   7      Score:34950
Group 7: Sequences:   8      Score:34414
Group 8: Sequences:   2      Score:29300
Group 9: Sequences:  10      Score:27647
Alignment Score 471014

CLUSTAL-Alignment file created  [SUP35_10seqs.aln]

3.39user 0.00system 0:03.39elapsed 100%CPU (0avgtext+0avgdata 5412maxresident)k
0inputs+96outputs (0major+837minor)pagefaults 0swaps
```

output - file in clustal format

In [4]:
```
! head -20 SUP35_10seqs.aln
```

```
CLUSTAL 2.1 multiple sequence alignment


SUP35_Sbou_unique28_CM003560        -----------ATGTCGGATTCAAAC-CAAGGCAACAATCAGCAAAACTA
SUP35_Scer_beer078_CM005938         -----------ATGTCGGATTCAAAC-CAAGGCAACAATCAGCAAAACTA
SUP35_Scer_74-D694_GCA_0015782      -----------ATGTCGGATTCAAAC-CAAGGCAACAATCAGCAAAACTA
SUP35_Spar_A12_Liti                 -----------ATGTCGGATTCAAAC-CAAGGTAACAATCAGCAAAGCTA
SUP35_Smik_IFO1815T_30              -----------ATGTCTGATTCAAAC-CAAGGTAATAATCAGCAAAACTA
SUP35_Sarb_H-6_chrXIII_CM00157      -----------ATGTCTGATCCAACT-AATGGTAATAATGAGCAGAGCTC
SUP35_Skud_IFO1802T_36              -----------ATGTCAGATCCAAAT-CAAGGTAACAATCAACAA-----
SUP35_Seub_CBS12357_chr_II_IV_      -----------ATGTCTGATCCAAAC-CAAGGTAACAATCAGCAAAACTA
SUP35_Kla_AB039749                  -----------ATGTCAGACCAA----CAAAATCAAGACCAAGGGCAAGG
SUP35_Agos_ATCC_10895_NM_21158      ATGTCGGAGGAAGATCAAATTCAATCGCAAGGCAACGACCAAGGCCAGTC
                                               *  **  *   *     *     *  *  *

SUP35_Sbou_unique28_CM003560        CCAGCAATACAGCCAGAACGGTAACCAACAACAAGGTAACAA------CA
SUP35_Scer_beer078_CM005938         CCAGCAATACAGCCAGAACGGTAACCAACAACAAGGTAACAA------CA
SUP35_Scer_74-D694_GCA_0015782      CCAGCAATACAGCCAGAACGGTAACCAACAACAAGGTAACAA------CA
SUP35_Spar_A12_Liti                 CCAGCAATACGGCCAAAACCCTAACCAACAACAAGGTAACAA------CA
SUP35_Smik_IFO1815T_30              CCAGCAATACAACCAAAACCCCAACCAACAGCAAGGTAACAA------TA
```

The same results from GUI

Specify output file

In [5]:
```
! clustalw -INFILE=SUP35_10seqs.fa -OUTFILE=SUP35_10seqs.clustalw
```

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: SUP35_Kla_AB039749                      2103 bp
Sequence 2: SUP35_Agos_ATCC_10895_NM_211584         2076 bp
Sequence 3: SUP35_Scer_74-D694_GCA_001578265.1      2058 bp
Sequence 4: SUP35_Spar_A12_Liti                     2046 bp
Sequence 5: SUP35_Smik_IFO1815T_30                  2046 bp
Sequence 6: SUP35_Skud_IFO1802T_36                  2037 bp
Sequence 7: SUP35_Sbou_unique28_CM003560            2058 bp
Sequence 8: SUP35_Scer_beer078_CM005938             2001 bp
Sequence 9: SUP35_Sarb_H-6_chrXIII_CM001575         2043 bp
Sequence 10: SUP35_Seub_CBS12357_chr_II_IV_DF968535  2025 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  70
Sequences (1:3) Aligned. Score:  71
Sequences (1:4) Aligned. Score:  71
Sequences (1:5) Aligned. Score:  71
Sequences (1:6) Aligned. Score:  71
Sequences (1:7) Aligned. Score:  71
Sequences (1:8) Aligned. Score:  69
Sequences (1:9) Aligned. Score:  62
Sequences (1:10) Aligned. Score:  73
Sequences (2:3) Aligned. Score:  66
Sequences (2:4) Aligned. Score:  69
Sequences (2:5) Aligned. Score:  69
Sequences (2:6) Aligned. Score:  69
Sequences (2:7) Aligned. Score:  66
Sequences (2:8) Aligned. Score:  70
Sequences (2:9) Aligned. Score:  69
Sequences (2:10) Aligned. Score:  69
Sequences (3:4) Aligned. Score:  93
Sequences (3:5) Aligned. Score:  88
Sequences (3:6) Aligned. Score:  85
Sequences (3:7) Aligned. Score:  99
Sequences (3:8) Aligned. Score:  99
Sequences (3:9) Aligned. Score:  86
Sequences (3:10) Aligned. Score:  84
Sequences (4:5) Aligned. Score:  89
Sequences (4:6) Aligned. Score:  86
Sequences (4:7) Aligned. Score:  93
Sequences (4:8) Aligned. Score:  93
Sequences (4:9) Aligned. Score:  87
Sequences (4:10) Aligned. Score:  84
Sequences (5:6) Aligned. Score:  85
Sequences (5:7) Aligned. Score:  88
Sequences (5:8) Aligned. Score:  88
Sequences (5:9) Aligned. Score:  86
Sequences (5:10) Aligned. Score:  84
Sequences (6:7) Aligned. Score:  85
Sequences (6:8) Aligned. Score:  85
Sequences (6:9) Aligned. Score:  86
Sequences (6:10) Aligned. Score:  85
Sequences (7:8) Aligned. Score:  99
Sequences (7:9) Aligned. Score:  86
Sequences (7:10) Aligned. Score:  84
Sequences (8:9) Aligned. Score:  85
Sequences (8:10) Aligned. Score:  84
Sequences (9:10) Aligned. Score:  86
Guide tree file created:   [SUP35_10seqs.dnd]

There are 9 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:37600
Group 2: Sequences:   3      Score:38275
Group 3: Sequences:   4      Score:36568
Group 4: Sequences:   5      Score:35737
Group 5: Sequences:   6      Score:35113
Group 6: Sequences:   7      Score:34950
Group 7: Sequences:   8      Score:34414
Group 8: Sequences:   2      Score:29300
Group 9: Sequences:  10      Score:27647
Alignment Score 471014

CLUSTAL-Alignment file created  [SUP35_10seqs.clustalw]
```

Save ClustalW results in fasta format

In [6]:
```
! clustalw -INFILE=SUP35_10seqs.fa -OUTPUT=FASTA -OUTFILE=SUP35_10seqs_clustalw.fa
```

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: SUP35_Kla_AB039749                      2103 bp
Sequence 2: SUP35_Agos_ATCC_10895_NM_211584         2076 bp
Sequence 3: SUP35_Scer_74-D694_GCA_001578265.1      2058 bp
Sequence 4: SUP35_Spar_A12_Liti                     2046 bp
Sequence 5: SUP35_Smik_IFO1815T_30                  2046 bp
Sequence 6: SUP35_Skud_IFO1802T_36                  2037 bp
Sequence 7: SUP35_Sbou_unique28_CM003560            2058 bp
Sequence 8: SUP35_Scer_beer078_CM005938             2001 bp
Sequence 9: SUP35_Sarb_H-6_chrXIII_CM001575         2043 bp
Sequence 10: SUP35_Seub_CBS12357_chr_II_IV_DF968535  2025 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  70
Sequences (1:3) Aligned. Score:  71
Sequences (1:4) Aligned. Score:  71
Sequences (1:5) Aligned. Score:  71
Sequences (1:6) Aligned. Score:  71
Sequences (1:7) Aligned. Score:  71
Sequences (1:8) Aligned. Score:  69
Sequences (1:9) Aligned. Score:  62
Sequences (1:10) Aligned. Score:  73
Sequences (2:3) Aligned. Score:  66
Sequences (2:4) Aligned. Score:  69
Sequences (2:5) Aligned. Score:  69
Sequences (2:6) Aligned. Score:  69
Sequences (2:7) Aligned. Score:  66
Sequences (2:8) Aligned. Score:  70
Sequences (2:9) Aligned. Score:  69
Sequences (2:10) Aligned. Score:  69
Sequences (3:4) Aligned. Score:  93
Sequences (3:5) Aligned. Score:  88
Sequences (3:6) Aligned. Score:  85
Sequences (3:7) Aligned. Score:  99
Sequences (3:8) Aligned. Score:  99
Sequences (3:9) Aligned. Score:  86
Sequences (3:10) Aligned. Score:  84
Sequences (4:5) Aligned. Score:  89
Sequences (4:6) Aligned. Score:  86
Sequences (4:7) Aligned. Score:  93
Sequences (4:8) Aligned. Score:  93
Sequences (4:9) Aligned. Score:  87
Sequences (4:10) Aligned. Score:  84
Sequences (5:6) Aligned. Score:  85
Sequences (5:7) Aligned. Score:  88
Sequences (5:8) Aligned. Score:  88
Sequences (5:9) Aligned. Score:  86
Sequences (5:10) Aligned. Score:  84
Sequences (6:7) Aligned. Score:  85
Sequences (6:8) Aligned. Score:  85
Sequences (6:9) Aligned. Score:  86
Sequences (6:10) Aligned. Score:  85
Sequences (7:8) Aligned. Score:  99
Sequences (7:9) Aligned. Score:  86
Sequences (7:10) Aligned. Score:  84
Sequences (8:9) Aligned. Score:  85
Sequences (8:10) Aligned. Score:  84
Sequences (9:10) Aligned. Score:  86
Guide tree file created:   [SUP35_10seqs.dnd]

There are 9 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:37600
Group 2: Sequences:   3      Score:38275
Group 3: Sequences:   4      Score:36568
Group 4: Sequences:   5      Score:35737
Group 5: Sequences:   6      Score:35113
Group 6: Sequences:   7      Score:34950
Group 7: Sequences:   8      Score:34414
Group 8: Sequences:   2      Score:29300
Group 9: Sequences:  10      Score:27647
Alignment Score 471014
firstres = 1 lastres = 2148
FASTA file created!

Fasta-Alignment file created    [SUP35_10seqs_clustalw.fa]
```

In [7]:
```
! head -30 SUP35_10seqs_clustalw.fa
```

```
>SUP35_Sbou_unique28_CM003560
-----------ATGTCGGATTCAAAC-CAAGGCAACAATC
AGCAAAACTACCAGCAATACAGCCAGAACGGTAACCAACA
ACAAGGTAACAA------CAGATACCAAGGTTATCAAGCT
TACAATGCTCAAGCCCAAC---CTGCAGGTGGGTACTACC
AAAATTACCAAGGTTATTCTGGGTACCAACAAGGTGGCTA
---TCAACAGTACAATCCCG---ACGCCGGTTACCAGCAA
CAGTATAATCCTCAAGGAGGCTATCAACAG------TACA
ATCCTCAAGGCGGTTATCAG--------CAGCAA----TT
CAATCCACAAGGTGGCCGTGGAAATTACAAAAACTTCAAC
TACAA-------TAACAATTTGCAAGGATATCAAGC---T
GGTTTC--------------CAACCACAGTCTCAAGGTAT
GTCTTTGAACGACTTTCAAAAGCAACAAAAGC------AG
GCCGCTCCCAAACCAAAGAAGACTTTGAAGCTTGTCTCCA
GTTCCGGTATCAAGTTGGCCAATGCTACCAAGAAGGTTGA
CACAAAACCTGCCGAATCTGATAAGAAAGAGGAAG---AG
AAGTCTGCTGAAACCAAAGAACCAACTAAAGAGCCAACAA
AGGTCGAAGAACCAGTTAAAAAGGAGGAGAAACCAGTCCA
GACTGAAGAA--AAGACGGAGGAAAAAT--CGGAACTTCC
AAAGGTAGAAGACC---TTAAAATCTCTGAATCAACACAT
A-ATACCA-ACAATGCCAATGTTACCAGTGCTGATGCCTT
GATCAAGGAACAGGAAGAAGAAGTGGATGACGAAGTTGTT
AACGATATGTTTGGTGGTAAAGATCACGTTTCTTTAATTT
TCATGGGTCATGTTGATGCCGGTAAATCTACTATGGGTGG
TAATCTACTATACTTGACTGGCTCTGTGGATAAGAGAACT
ATTGAGAAATATGAAAGAGAAGCCAAGGATGCAGGCAGAC
AAGGTTGGTACTTGTCATGGGTCATGGATACCAACAAAGA
AGAAAGAAATGATGGTAAGACTATCGAAGTTGGTAAGGCC
TACTTTGAAACTGAAAAAAGGCGTTATACCATATTGGATG
CTCCTGGTCATAAAATGTACGTTTCCGAGATGATCGGTGG
```

In [8]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_clustalw.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Sbou_unique28_CM003560
  2148
SUP35_Scer_beer078_CM005938
  2148
SUP35_Scer_74-D694_GCA_001578265.1
  2148
SUP35_Spar_A12_Liti
  2148
SUP35_Smik_IFO1815T_30
  2148
SUP35_Sarb_H-6_chrXIII_CM001575
  2148
SUP35_Skud_IFO1802T_36
  2148
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2148
SUP35_Kla_AB039749
  2148
SUP35_Agos_ATCC_10895_NM_211584
  2148
```

Muscle

In [9]:
```
! time muscle -in SUP35_10seqs.fa -out SUP35_10seqs_muscle.fa
```

```
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

SUP35_10seqs 10 seqs, lengths min 2001, max 2103, avg 2049
00:00:00    16 MB(-2%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    16 MB(-2%)  Iter   1  100.00%  K-mer dist pass 2
00:00:00    37 MB(-5%)  Iter   1  100.00%  Align node
00:00:00    37 MB(-5%)  Iter   1  100.00%  Root alignment
00:00:01    37 MB(-5%)  Iter   2  100.00%  Refine tree
00:00:01    37 MB(-5%)  Iter   2  100.00%  Root alignment
00:00:01    37 MB(-5%)  Iter   2  100.00%  Root alignment
00:00:02    37 MB(-5%)  Iter   3  100.00%  Refine biparts
00:00:03    37 MB(-5%)  Iter   4  100.00%  Refine biparts
00:00:03    37 MB(-5%)  Iter   5  100.00%  Refine biparts
00:00:03    37 MB(-5%)  Iter   5  100.00%  Refine biparts
3.32user 0.00system 0:03.33elapsed 100%CPU (0avgtext+0avgdata 25156maxresident)k
0inputs+48outputs (0major+6734minor)pagefaults 0swaps
```

In [10]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_muscle.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Agos_ATCC_10895_NM_211584
  2275
SUP35_Kla_AB039749
  2275
SUP35_Skud_IFO1802T_36
  2275
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2275
SUP35_Sarb_H-6_chrXIII_CM001575
  2275
SUP35_Smik_IFO1815T_30
  2275
SUP35_Spar_A12_Liti
  2275
SUP35_Scer_74-D694_GCA_001578265.1
  2275
SUP35_Sbou_unique28_CM003560
  2275
SUP35_Scer_beer078_CM005938
  2275
```

MAFFT

`--auto` : search optimal parameters depending on the number of input sequences (search trade off between speed and used resources)

In [11]:
```
! time mafft --auto SUP35_10seqs.fa >SUP35_10seqs_mafft.fa
```

```
outputhat23=16
treein = 0
compacttree = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
All-to-all alignment.
tbfast-pair (nuc) Version 7.453
alg=L, model=DNA200 (2), 2.00 (6.00), -0.10 (-0.30), noshift, amax=0.0
0 thread(s)

outputhat23=16
Loading 'hat3.seed' ...
done.
Writing hat3 for iterative refinement
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00
tbutree = 1, compacttree = 0
Constructing a UPGMA tree ...
    0 / 10
done.

Progressive alignment ...
STEP     8 /9
Reallocating..done. *alloclen = 5227
STEP     9 /9
done.
tbfast (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
1 thread(s)

minimumweight = 0.000010
autosubalignment = 0.000000
nthread = 0
randomseed = 0
blosum 62 / kimura 200
poffset = 0
niter = 16
sueff_global = 0.100000
nadd = 16
Loading 'hat3' ... done.
generating a scoring matrix for nucleotide (dist=200) ... done

    0 / 10
Segment   1/  1    1-2219
STEP 003-003-1  identical.    identical.    identical.    accepted. accepted. identical.    identica
l.    identical.
Converged.

done
dvtditr (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 L-INS-i (Probably most accurate, very slow)
 Iterative refinement method (<16) with LOCAL pairwise alignment information

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.

2.68user 0.09system 0:02.70elapsed 102%CPU (0avgtext+0avgdata 71160maxresident)k
0inputs+280outputs (0major+56394minor)pagefaults 0swaps
```

In [12]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_mafft.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749
  2166
SUP35_Agos_ATCC_10895_NM_211584
  2166
SUP35_Scer_74-D694_GCA_001578265.1
  2166
SUP35_Spar_A12_Liti
  2166
SUP35_Smik_IFO1815T_30
  2166
SUP35_Skud_IFO1802T_36
  2166
SUP35_Sbou_unique28_CM003560
  2166
SUP35_Scer_beer078_CM005938
  2166
SUP35_Sarb_H-6_chrXIII_CM001575
  2166
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2166
```

Kalign

In [13]:
```
! time kalign <SUP35_10seqs.fa >SUP35_10seqs_kalign.fa
```

```
Kalign version 2.04, Copyright (C) 2004, 2005, 2006 Timo Lassmann

        Kalign is free software. You can redistribute it and/or modify
        it under the terms of the GNU General Public License as
        published by the Free Software Foundation.

reading from STDIN: found 10 sequences
Aligning 10 RNA/DNA sequences with these parameters:
        217.00000000    gap open penalty
        39.40000153     gap extension
        292.60000610    terminal gap penalty
        283.00000000    bonus
Alignment will be written to stdout.

Distance Calculation:
     100 percent done
Alignment:
       0 percent doneSaving mem...
      10 percent doneSaving mem...
      20 percent doneSaving mem...
      30 percent doneSaving mem...
      40 percent doneSaving mem...
      50 percent doneSaving mem...
      60 percent doneSaving mem...
      70 percent doneSaving mem...
      80 percent doneSaving mem...
     100 percent done
0.33user 0.00system 0:00.34elapsed 99%CPU (0avgtext+0avgdata 2932maxresident)k
0inputs+48outputs (0major+656minor)pagefaults 0swaps
```

In [14]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_kalign.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749
  2155
SUP35_Agos_ATCC_10895_NM_211584
  2155
SUP35_Scer_74-D694_GCA_001578265.1
  2155
SUP35_Spar_A12_Liti
  2155
SUP35_Smik_IFO1815T_30
  2155
SUP35_Skud_IFO1802T_36
  2155
SUP35_Sbou_unique28_CM003560
  2155
SUP35_Scer_beer078_CM005938
  2155
SUP35_Sarb_H-6_chrXIII_CM001575
  2155
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2155
```

t-coffee

Slow and requires many resources! Be careful!

In [15]:
```
! time t_coffee -infile=SUP35_10seqs.fa -outfile=SUP35_10seqs_tcoffee.fa -output=fasta_aln > t_coffee_10.log
```

```
PROGRAM: T-COFFEE Version_11.00.8cbe486 (2014-08-12 22:05:29 - Revision 8cbe486 - Build 477)
-full_log       S       [0]
-genepred_score S       [0]     nsd
-run_name       S       [0]
-mem_mode       S       [0]     mem
-extend         D       [1]     1
-extend_mode    S       [0]     very_fast_triplet
-max_n_pair     D       [0]     10
-seq_name_for_quadruplet        S       [0]     all
-compact        S       [0]     default
-clean          S       [0]     no
-do_self        FL      [0]     0
-do_normalise   D       [0]     1000
-template_file  S       [0]
-setenv         S       [0]     0
-template_mode  S       [0]
-flip           D       [0]     0
-remove_template_file   D       [0]     0
-profile_template_file  S       [0]
-in             S       [0]
-seq            S       [0]
-aln            S       [0]
-method_limits  S       [0]
-method         S       [0]
-lib            S       [0]
-profile        S       [0]
-profile1       S       [0]
-profile2       S       [0]
-pdb            S       [0]
-relax_lib      D       [0]     1
-filter_lib     D       [0]     0
-shrink_lib     D       [0]     0
-out_lib        W_F     [0]     no
-out_lib_mode   S       [0]     primary
-lib_only       D       [0]     0
-outseqweight   W_F     [0]     no
-dpa            FL      [0]     0
-seq_source     S       [0]     ANY
-cosmetic_penalty       D       [0]     0
-gapopen        D       [0]     0
-gapext         D       [0]     0
-fgapopen       D       [0]     0
-fgapext        D       [0]     0
-nomatch        D       [0]     0
-newtree        W_F     [0]     default
-tree           W_F     [0]     NO
-usetree        R_F     [0]
-tree_mode      S       [0]     nj
-distance_matrix_mode   S       [0]     ktup
-distance_matrix_sim_mode       S       [0]     idmat_sim1
-quicktree      FL      [0]     0
-outfile        W_F     [1]     SUP35_10seqs_tcoffee.fa
-maximise       FL      [1]     1
-output         S       [1]     fasta_aln
-len            D       [0]     0
-infile         R_F     [1]     SUP35_10seqs.fa
-matrix         S       [0]     default
-tg_mode        D       [0]     1
-profile_mode   S       [0]     cw_profile_profile
-profile_comparison     S       [0]     profile
-dp_mode        S       [0]     linked_pair_wise
-ktuple         D       [0]     1
-ndiag          D       [0]     0
-diag_threshold D       [0]     0
-diag_mode      D       [0]     0
-sim_matrix     S       [0]     vasiliky
-transform      S       [0]
-extend_seq     FL      [0]     0
-outorder       S       [0]     input
-inorder        S       [0]     aligned
-seqnos         S       [0]     off
-case           S       [0]     keep
-cpu            D       [0]     0
-ulimit         D       [0]     -1
-maxnseq        D       [0]     -1
-maxlen         D       [0]     -1
-sample_dp      D       [0]     0
-weight         S       [0]     default
-seq_weight     S       [0]     no
-align          FL      [1]     1
-mocca          FL      [0]     0
-domain         FL      [0]     0
-start          D       [0]     0
-len            D       [0]     0
-scale          D       [0]     0
-mocca_interactive      FL      [0]     0
-method_evaluate_mode   S       [0]     default
-evaluate_mode  S       [0]     triplet
-get_type       FL      [0]     0
-clean_aln      D       [0]     0
-clean_threshold        D       [1]     1
-clean_iteration        D       [1]     1
-clean_evaluate_mode    S       [0]     t_coffee_fast
-extend_matrix  FL      [0]     0
-prot_min_sim   D       [40]    40
-prot_max_sim   D       [90]    90
-prot_min_cov   D       [40]    40
-pdb_type       S       [0]     d
-pdb_min_sim    D       [35]    35
-pdb_max_sim    D       [100]   100
-pdb_min_cov    D       [50]    50
-pdb_blast_server       W_F     [0]     EBI
-blast          W_F     [0]
-blast_server   W_F     [0]     EBI
-pdb_db         W_F     [0]     pdb
-protein_db     W_F     [0]     uniprot
-method_log     W_F     [0]     no
-struc_to_use   S       [0]
-cache          W_F     [0]     use
-align_pdb_param_file   W_F     [0]     no
-align_pdb_hasch_mode   W_F     [0]     hasch_ca_trace_bubble
-external_aligner       S       [0]     NO
-msa_mode       S       [0]     tree
-master         S       [0]     no
-blast_nseq     D       [0]     0
-lalign_n_top   D       [0]     10
-iterate        D       [0]     0
-trim           D       [0]     0
-split          D       [0]     0
-trimfile       S       [0]     default
-split          D       [0]     0
-split_nseq_thres       D       [0]     0
-split_score_thres      D       [0]     0
-check_pdb_status       D       [0]     0
-clean_seq_name D       [0]     0
-seq_to_keep    S       [0]
-dpa_master_aln S       [0]
-dpa_maxnseq    D       [0]     0
-dpa_min_score1 D       [0]
-dpa_min_score2 D       [0]
-dpa_keep_tmpfile       FL      [0]     0
-dpa_debug      D       [0]     0
-multi_core     S       [0]     templates_jobs_relax_msa_evaluate
-n_core         D       [0]     0
-max_n_proc     D       [0]     0
-lib_list       S       [0]
-prune_lib_mode S       [0]     5
-tip            S       [0]     none
-rna_lib        S       [0]
-no_warning     D       [0]     0
-run_local_script       D       [0]     0
-plugins        S       [0]     default
-proxy          S       [0]     unset
-email          S       [0]
-clean_overaln  D       [0]     0
-overaln_param  S       [0]
-overaln_mode   S       [0]
-overaln_model  S       [0]
-overaln_threshold      D       [0]     0
-overaln_target D       [0]     0
-overaln_P1     D       [0]     0
-overaln_P2     D       [0]     0
-overaln_P3     D       [0]     0
-overaln_P4     D       [0]     0
-exon_boundaries        S       [0]
-dump           S       [0]     no
-display        D       [0]     100

INPUT FILES
        Input File (S) SUP35_10seqs.fa  Format fasta_seq
        Input File (M) proba_pair

Identify Master Sequences [no]:

Master Sequences Identified
INPUT SEQUENCES: 10 SEQUENCES  [DNA]
  Input File SUP35_10seqs.fa                        Seq SUP35_Agos_ATCC_10895_NM_211584        Length
2076 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Kla_AB039749                     Length
2103 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Sarb_H-6_chrXIII_CM001575        Length
2043 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Sbou_unique28_CM003560           Length
2058 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Scer_74-D694_GCA_001578265.1     Length
2058 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Scer_beer078_CM005938            Length
2001 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Seub_CBS12357_chr_II_IV_DF968535 Length
2025 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Skud_IFO1802T_36                 Length
2037 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Smik_IFO1815T_30                 Length
2046 type DNA Struct Unchecked
  Input File SUP35_10seqs.fa                        Seq SUP35_Spar_A12_Liti                    Length
2046 type DNA Struct Unchecked

        Multi Core Mode: 8 processors:

        --- Process Method/Library/Aln SSUP35_10seqs.fa
        --- Process Method/Library/Aln Mproba_pair
        xxx Retrieved SSUP35_10seqs.fa
        xxx Retrieved Mproba_pair

        All Methods Retrieved

MANUAL PENALTIES: gapopen=0 gapext=0

        Library Total Size: [271542]

Library Relaxation: Multi_proc [8]

!               [Relax Library][TOT=    1][100 %][ ELAPSED  TIME:    0 sec.]
Relaxation Summary: [271542]--->[231098]



UN-WEIGHTED MODE: EVERY SEQUENCE WEIGHTS 1

MAKE GUIDE TREE
        [MODE=nj][DONE]

PROGRESSIVE_ALIGNMENT [Tree Based]
Group    1: SUP35_Agos_ATCC_10895_NM_211584
Group    2: SUP35_Kla_AB039749
Group    3: SUP35_Sarb_H-6_chrXIII_CM001575
Group    4: SUP35_Sbou_unique28_CM003560
Group    5: SUP35_Scer_74-D694_GCA_001578265.1
Group    6: SUP35_Scer_beer078_CM005938
Group    7: SUP35_Seub_CBS12357_chr_II_IV_DF968535
Group    8: SUP35_Skud_IFO1802T_36
Group    9: SUP35_Smik_IFO1815T_30
Group   10: SUP35_Spar_A12_Liti


        Group   11: [Group    7 (   1 seq)] with [Group    3 (   1 seq)]-->[Len= 2052][PID:23209]
        Group   12: [Group    5 (   1 seq)] with [Group    4 (   1 seq)]-->[Len= 2058][PID:23210]
        Group   13: [Group    6 (   1 seq)] with [Group   12 (   2 seq)]-->[Len= 2058][PID:23210]
        Group   14: [Group   10 (   1 seq)] with [Group   13 (   3 seq)]-->[Len= 2058][PID:23210]
        Group   15: [Group    9 (   1 seq)] with [Group   14 (   4 seq)]-->[Len= 2058][PID:23210]
        Group   16: [Group   15 (   5 seq)] with [Group   11 (   2 seq)]-->[Len= 2064][PID:23178][For
ked]
        Group   17: [Group   16 (   7 seq)] with [Group    8 (   1 seq)]-->[Len= 2064][PID:23178]
        Group   18: [Group    2 (   1 seq)] with [Group   17 (   8 seq)]-->[Len= 2141][PID:23178]
        Group   19: [Group    1 (   1 seq)] with [Group   18 (   9 seq)]-->[Len= 2210][PID:23178]


!               [Final Evaluation][TOT=  276][100 %][ ELAPSED  TIME:    0 sec.]



OUTPUT RESULTS
        #### File Type= GUIDE_TREE      Format= newick          Name= SUP35_10seqs.dnd
        #### File Type= MSA             Format= fasta_aln       Name= SUP35_10seqs_tcoffee.fa

# Command Line: /home/isemenov/anaconda3/lib/t_coffee-11.0.8/bin/t_coffee -infile SUP35_10seqs.fa -ou
tfile SUP35_10seqs_tcoffee.fa -output fasta_aln  [PROGRAM:T-COFFEE]
# T-COFFEE Memory Usage: Current= 37.060 Mb, Max= 44.536 Mb
# Results Produced with T-COFFEE Version_11.00.8cbe486 (2014-08-12 22:05:29 - Revision 8cbe486 - Buil
d 477)
# T-COFFEE is available from http://www.tcoffee.org
# Register on: https://groups.google.com/group/tcoffee/
278.58user 5.67system 0:43.26elapsed 656%CPU (0avgtext+0avgdata 675740maxresident)k
13760inputs+27840outputs (73major+1334525minor)pagefaults 0swaps
```

In [16]:
```python
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_tcoffee.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749
  2210
SUP35_Agos_ATCC_10895_NM_211584
  2210
SUP35_Scer_74-D694_GCA_001578265.1
  2210
SUP35_Spar_A12_Liti
  2210
SUP35_Smik_IFO1815T_30
  2210
SUP35_Skud_IFO1802T_36
  2210
SUP35_Sbou_unique28_CM003560
  2210
SUP35_Scer_beer078_CM005938
  2210
SUP35_Sarb_H-6_chrXIII_CM001575
  2210
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2210
```

Prank

In [17]:
```
! time prank -d=SUP35_10seqs.fa -o=SUP35_10seqs_prank.fa
```

```
-----------------
 PRANK v.170427:
-----------------

Input for the analysis
 - aligning sequences in 'SUP35_10seqs.fa'
 - using inferred alignment guide tree
 - option '+F' is not used; it can be enabled with '+F'
 - external tools available:
    MAFFT for initial alignment
    Exonerate for alignment anchoring

Correcting (arbitrarily) for multifurcating nodes.
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 1.

Alignment score: 4546
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 2.

Alignment score: 4981
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 3.

Alignment score: 4981
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 4.

Alignment score: 4981
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 5.

Alignment score: 4981


Writing
 - alignment to 'SUP35_10seqs_prank.fa.best.fas'

Analysis done. Total time 7s

6.78user 0.11system 0:06.92elapsed 99%CPU (0avgtext+0avgdata 32636maxresident)k
23016inputs+472outputs (131major+60430minor)pagefaults 0swaps
```

In [18]:
```python
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_prank.fa.best.fas", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Agos_ATCC_10895_NM_211584
  2366
SUP35_Sarb_H-6_chrXIII_CM001575
  2366
SUP35_Seub_CBS12357_chr_II_IV_DF968535
  2366
SUP35_Skud_IFO1802T_36
  2366
SUP35_Scer_74-D694_GCA_001578265.1
  2366
SUP35_Sbou_unique28_CM003560
  2366
SUP35_Scer_beer078_CM005938
  2366
SUP35_Spar_A12_Liti
  2366
SUP35_Smik_IFO1815T_30
  2366
SUP35_Kla_AB039749
  2366
```

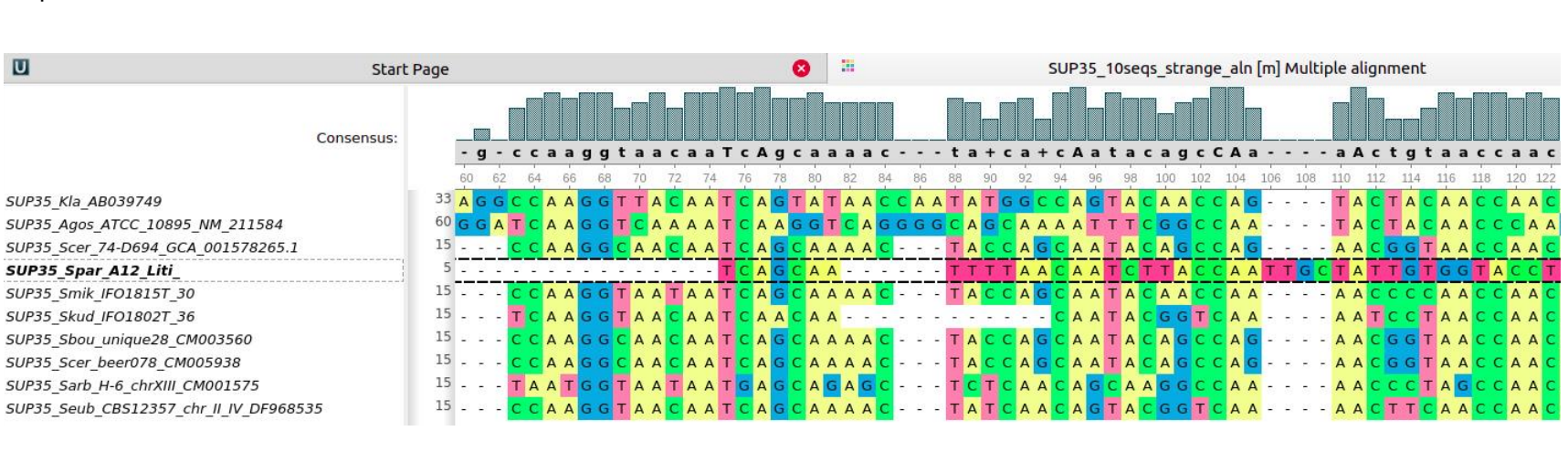
Compare algorithms on 10 sequences

algorithm	alignment length (nt)	time (seconds)
ClustalW	2148	3.30
MUSCLE	2275	3.02
MAFFT	2166	2.49
Kalign	2155	0.30
T-Coffee	2210	262.50
prank	2366	6.11

Conclusion: The best algorithm here is **Prank**. It has comparable speed with other algorithms and the highest length of alignment among them.

Strange file `SUP35_10seqs_strange_aln.fa`

As we can see there is one strange sequence in the file (highlighted) which does not match well with others.
It is connceted with that fact that this seqience is a reverse-complement. So, we should obtain it's reverse-complement and then realign sequences.



After reverse-complemet operation:

Launch different alignment algorithms on the file SUP35_250seqs.fa

`SUP35_250seqs.fa` contains 250 sequences of the gene SUP35 from different S. cerevisiae strains

ClustalW

In [19]:
```
! time clustalw -INFILE=SUP35_250seqs.fa -OUTPUT=FASTA -OUTFILE=SUP35_250seqs_clustalw.fa > clustalw_250.log
```

```
1541.83user 0.18system 25:43.73elapsed 99%CPU (0avgtext+0avgdata 13596maxresident)k
4200inputs+4192outputs (20major+3042minor)pagefaults 0swaps
```

In [20]:
```python
record_dict = IO.to_dict(IO.parse("SUP35_250seqs_clustalw.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
    break
```

```
SUP35_Scer_T.52_2H_CP007848
  2179
```

Muscle

In [21]:
```
! time muscle -in SUP35_250seqs.fa -out SUP35_250seqs_muscle.fa
```

```
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

SUP35_250seqs 250 seqs, lengths min 1769, max 2103, avg 2054
00:00:00    19 MB(-2%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    19 MB(-2%)  Iter   1  100.00%  K-mer dist pass 2
00:00:14  332 MB(-43%)  Iter   1  100.00%  Align node
00:00:14  333 MB(-43%)  Iter   1  100.00%  Root alignment
00:00:19  333 MB(-43%)  Iter   2  100.00%  Refine tree
00:00:19  333 MB(-43%)  Iter   2  100.00%  Root alignment
00:00:19  333 MB(-43%)  Iter   2  100.00%  Root alignment
00:01:02  333 MB(-43%)  Iter   3  100.00%  Refine biparts
00:01:46  333 MB(-43%)  Iter   4  100.00%  Refine biparts
00:01:46  333 MB(-43%)  Iter   5  100.00%  Refine biparts
00:01:46  333 MB(-43%)  Iter   5  100.00%  Refine biparts
105.41user 0.18system 1:45.85elapsed 99%CPU (0avgtext+0avgdata 312612maxresident)k
1152inputs+1168outputs (9major+80479minor)pagefaults 0swaps
```

In [22]:
```python
record_dict = IO.to_dict(IO.parse("SUP35_250seqs_muscle.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
    break
```

```
SUP35_Agos_ATCC_10895_NM_211584
  2313
```

MAFFT

In [23]:
```
! time mafft --auto SUP35_250seqs.fa >SUP35_250seqs_mafft.fa
```

```
nthread = 0
nthreadpair = 0
nthreadtb = 0
ppenalty_ex = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00



Making a distance matrix ..

There are 12 ambiguous characters.
  201 / 250
done.

Constructing a UPGMA tree (efffree=0) ...
  240 / 250
done.

Progressive alignment 1/2...
STEP   247 / 249
Reallocating..done. *alloclen = 5227
STEP   249 / 249
done.

Making a distance matrix from msa..
  200 / 250
done.

Constructing a UPGMA tree (efffree=1) ...
  240 / 250
done.

Progressive alignment 2/2...
STEP   247 / 249
Reallocating..done. *alloclen = 5227
STEP   249 / 249
done.

disttbfast (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)

generating a scoring matrix for nucleotide (dist=200) ... done
dndpre (nuc) Version 7.453
alg=X, model=DNA200 (2), 1.53 (4.59), 0.37 (1.11), noshift, amax=0.0
0 thread(s)

minimumweight = 0.000010
autosubalignment = 0.000000
nthread = 0
randomseed = 0
blosum 62 / kimura 200
poffset = 0
niter = 2
sueff_global = 0.100000
nadd = 2
generating a scoring matrix for nucleotide (dist=200) ... done

  240 / 250
Segment   1/ 20    1- 145
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.
Converged.

Segment   2/ 20  145- 248
STEP 002-014-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    rejected. identical.    identica
l.    identical.    identical.    identical.    identical.    identical.    identical.
Converged.

Segment   3/ 20  248- 299
Segment   4/ 20  299- 367.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.
STEP 002-013-0  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.
Converged.

Segment   5/ 20  361- 400
Segment   6/ 20  400- 462.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.
STEP 002-003-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.
Converged.

Segment   7/ 20  462- 608
Segment   8/ 20  608- 714.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.    identical.    identical.    identical.    identical.    identical.    identic
al.    identical.
Segment   9/ 20  714- 807.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.    identical.    identical.
Segment  10/ 20  807- 911.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    rejected. identical.    identical.    identical.    identical.    identica
l.    identical.    identical.
Segment  11/ 20  911-1045.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.    identical.    identical.
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.
Converged.

Segment  12/ 20 1045-1196
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.
Converged.

Segment  13/ 20 1196-1347
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.
Converged.

Segment  14/ 20 1347-1498
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.
Converged.

Segment  15/ 20 1498-1649
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.
Converged.

Segment  16/ 20 1649-1800
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.
Converged.

Segment  17/ 20 1800-1951
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.    identical.
Converged.

Segment  18/ 20 1951-2102
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.
Converged.

Segment  19/ 20 2102-2245
STEP 002-248-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.    identical.
Converged.

Segment  20/ 20 2245-2323
done 002-001-1  identical.    identical.    identical.    identical.    identical.    identical.    i
dentical.    identical.    identical.    identical.    identical.    identical.    identical.    iden
tical.
dvtditr (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-i (Standard)
 Iterative refinement method (max. 2 iterations)

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.

17.33user 0.15system 0:17.49elapsed 99%CPU (0avgtext+0avgdata 60600maxresident)k
11584inputs+5040outputs (52major+56496minor)pagefaults 0swaps
```

In [24]:
```python
record_dict = IO.to_dict(IO.parse("SUP35_250seqs_mafft.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
    break
```

```
SUP35_Kla_AB039749
  2322
```

Kalign

In [25]:
```
! time kalign <SUP35_250seqs.fa >SUP35_250seqs_kalign.fa
```

```
Kalign version 2.04, Copyright (C) 2004, 2005, 2006 Timo Lassmann

        Kalign is free software. You can redistribute it and/or modify
        it under the terms of the GNU General Public License as
        published by the Free Software Foundation.

reading from STDIN: found 250 sequences
Aligning 250 RNA/DNA sequences with these parameters:
        217.00000000    gap open penalty
        39.40000153     gap extension
        292.60000610    terminal gap penalty
        283.00000000    bonus
Alignment will be written to stdout.

Distance Calculation:
     100 percent done
Alignment:
       0 percent doneSaving mem...
       0 percent doneSaving mem...
       1 percent doneSaving mem...
       1 percent doneSaving mem...
       2 percent doneSaving mem...
       2 percent doneSaving mem...
       2 percent doneSaving mem...
       3 percent doneSaving mem...
       3 percent doneSaving mem...
       4 percent doneSaving mem...
       4 percent doneSaving mem...
       4 percent doneSaving mem...
       5 percent doneSaving mem...
       5 percent doneSaving mem...
       6 percent doneSaving mem...
       6 percent doneSaving mem...
       6 percent doneSaving mem...
       7 percent doneSaving mem...
       7 percent doneSaving mem...
       8 percent doneSaving mem...
       8 percent doneSaving mem...
       8 percent doneSaving mem...
       9 percent doneSaving mem...
       9 percent doneSaving mem...
      10 percent doneSaving mem...
      10 percent doneSaving mem...
      10 percent doneSaving mem...
      11 percent doneSaving mem...
      11 percent doneSaving mem...
      12 percent doneSaving mem...
      12 percent doneSaving mem...
      12 percent doneSaving mem...
      13 percent doneSaving mem...
      13 percent doneSaving mem...
      14 percent doneSaving mem...
      14 percent doneSaving mem...
      14 percent doneSaving mem...
      15 percent doneSaving mem...
      15 percent doneSaving mem...
      16 percent doneSaving mem...
      16 percent doneSaving mem...
      16 percent doneSaving mem...
      17 percent doneSaving mem...
      17 percent doneSaving mem...
      18 percent doneSaving mem...
      18 percent doneSaving mem...
      18 percent doneSaving mem...
      19 percent doneSaving mem...
      19 percent doneSaving mem...
      20 percent doneSaving mem...
      20 percent doneSaving mem...
      20 percent doneSaving mem...
      21 percent doneSaving mem...
      21 percent doneSaving mem...
      22 percent doneSaving mem...
      22 percent doneSaving mem...
      22 percent doneSaving mem...
      23 percent doneSaving mem...
      23 percent doneSaving mem...
      24 percent doneSaving mem...
      24 percent doneSaving mem...
      24 percent doneSaving mem...
      25 percent doneSaving mem...
      25 percent doneSaving mem...
      26 percent doneSaving mem...
      26 percent doneSaving mem...
      26 percent doneSaving mem...
      27 percent doneSaving mem...
      27 percent doneSaving mem...
      28 percent doneSaving mem...
      28 percent doneSaving mem...
      28 percent doneSaving mem...
      29 percent doneSaving mem...
      29 percent doneSaving mem...
      30 percent doneSaving mem...
      30 percent doneSaving mem...
      30 percent doneSaving mem...
      31 percent doneSaving mem...
      31 percent doneSaving mem...
      32 percent doneSaving mem...
      32 percent doneSaving mem...
      32 percent doneSaving mem...
      33 percent doneSaving mem...
      33 percent doneSaving mem...
      34 percent doneSaving mem...
      34 percent doneSaving mem...
      34 percent doneSaving mem...
      35 percent doneSaving mem...
      35 percent doneSaving mem...
      36 percent doneSaving mem...
      36 percent doneSaving mem...
      36 percent doneSaving mem...
      37 percent doneSaving mem...
      37 percent doneSaving mem...
      38 percent doneSaving mem...
      38 percent doneSaving mem...
      38 percent doneSaving mem...
      39 percent doneSaving mem...
      39 percent doneSaving mem...
      40 percent doneSaving mem...
      40 percent doneSaving mem...
      40 percent doneSaving mem...
      41 percent doneSaving mem...
      41 percent doneSaving mem...
      42 percent doneSaving mem...
      42 percent doneSaving mem...
      42 percent doneSaving mem...
      43 percent doneSaving mem...
      43 percent doneSaving mem...
      44 percent doneSaving mem...
      44 percent doneSaving mem...
      44 percent doneSaving mem...
      45 percent doneSaving mem...
      45 percent doneSaving mem...
      46 percent doneSaving mem...
      46 percent doneSaving mem...
      46 percent doneSaving mem...
      47 percent doneSaving mem...
      47 percent doneSaving mem...
      48 percent doneSaving mem...
      48 percent doneSaving mem...
      48 percent doneSaving mem...
      49 percent doneSaving mem...
      49 percent doneSaving mem...
      50 percent doneSaving mem...
      50 percent doneSaving mem...
      50 percent doneSaving mem...
      51 percent doneSaving mem...
      51 percent doneSaving mem...
      52 percent doneSaving mem...
      52 percent doneSaving mem...
      52 percent doneSaving mem...
      53 percent doneSaving mem...
      53 percent doneSaving mem...
      54 percent doneSaving mem...
      54 percent doneSaving mem...
      54 percent doneSaving mem...
      55 percent doneSaving mem...
      55 percent doneSaving mem...
      56 percent doneSaving mem...
      56 percent doneSaving mem...
      56 percent doneSaving mem...
      57 percent doneSaving mem...
      57 percent doneSaving mem...
      58 percent doneSaving mem...
      58 percent doneSaving mem...
      58 percent doneSaving mem...
      59 percent doneSaving mem...
      59 percent doneSaving mem...
      60 percent doneSaving mem...
      60 percent doneSaving mem...
      60 percent doneSaving mem...
      61 percent doneSaving mem...
      61 percent doneSaving mem...
      62 percent doneSaving mem...
      62 percent doneSaving mem...
      62 percent doneSaving mem...
      63 percent doneSaving mem...
      63 percent doneSaving mem...
      64 percent doneSaving mem...
      64 percent doneSaving mem...
      64 percent doneSaving mem...
      65 percent doneSaving mem...
      65 percent doneSaving mem...
      66 percent doneSaving mem...
      66 percent doneSaving mem...
      66 percent doneSaving mem...
      67 percent doneSaving mem...
      67 percent doneSaving mem...
      68 percent doneSaving mem...
      68 percent doneSaving mem...
      68 percent doneSaving mem...
      69 percent doneSaving mem...
      69 percent doneSaving mem...
      70 percent doneSaving mem...
      70 percent doneSaving mem...
      70 percent doneSaving mem...
      71 percent doneSaving mem...
      71 percent doneSaving mem...
      72 percent doneSaving mem...
      72 percent doneSaving mem...
      72 percent doneSaving mem...
      73 percent doneSaving mem...
      73 percent doneSaving mem...
      74 percent doneSaving mem...
      74 percent doneSaving mem...
      74 percent doneSaving mem...
      75 percent doneSaving mem...
      75 percent doneSaving mem...
      76 percent doneSaving mem...
      76 percent doneSaving mem...
      76 percent doneSaving mem...
      77 percent doneSaving mem...
      77 percent doneSaving mem...
      78 percent doneSaving mem...
      78 percent doneSaving mem...
      78 percent doneSaving mem...
      79 percent doneSaving mem...
      79 percent doneSaving mem...
      80 percent doneSaving mem...
      80 percent doneSaving mem...
      80 percent doneSaving mem...
      81 percent doneSaving mem...
      81 percent doneSaving mem...
      82 percent doneSaving mem...
      82 percent doneSaving mem...
      82 percent doneSaving mem...
      83 percent doneSaving mem...
      83 percent doneSaving mem...
      84 percent doneSaving mem...
      84 percent doneSaving mem...
      84 percent doneSaving mem...
      85 percent doneSaving mem...
      85 percent doneSaving mem...
      86 percent doneSaving mem...
      86 percent doneSaving mem...
      86 percent doneSaving mem...
      87 percent doneSaving mem...
      87 percent doneSaving mem...
      88 percent doneSaving mem...
      88 percent doneSaving mem...
      88 percent doneSaving mem...
      89 percent doneSaving mem...
      89 percent doneSaving mem...
      90 percent doneSaving mem...
      90 percent doneSaving mem...
      90 percent doneSaving mem...
      91 percent doneSaving mem...
      91 percent doneSaving mem...
      92 percent doneSaving mem...
      92 percent doneSaving mem...
      92 percent doneSaving mem...
      93 percent doneSaving mem...
      93 percent doneSaving mem...
      94 percent doneSaving mem...
      94 percent doneSaving mem...
      94 percent doneSaving mem...
      95 percent doneSaving mem...
      95 percent doneSaving mem...
      96 percent doneSaving mem...
      96 percent doneSaving mem...
      96 percent doneSaving mem...
      97 percent doneSaving mem...
      97 percent doneSaving mem...
      98 percent doneSaving mem...
      98 percent doneSaving mem...
      98 percent doneSaving mem...
      99 percent doneSaving mem...
      99 percent doneSaving mem...
     100 percent done
16.93user 0.08system 0:17.05elapsed 99%CPU (0avgtext+0avgdata 8092maxresident)k
456inputs+1120outputs (4major+1802minor)pagefaults 0swaps
```

In [26]:
```
record_dict = IO.to_dict(IO.parse("SUP35_250seqs_kalign.fa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
    break
```

```
SUP35_Kla_AB039749
  2214
```

t-coffee

It is too slow to launch it on 250 sequences.

In [27]:
```
# ! time t_coffee -infile=SUP35_250seqs.fa -outfile=SUP35_250seqs_tcoffee.fa -output=fasta_aln
```

Prank

In [28]:
```
! time prank -d=SUP35_250seqs.fa -o=SUP35_250seqs_prank.fa
```

```
-----------------
 PRANK v.170427:
-----------------

Input for the analysis
 - aligning sequences in 'SUP35_250seqs.fa'
 - using inferred alignment guide tree
 - option '+F' is not used; it can be enabled with '+F'
 - external tools available:
    MAFFT for initial alignment
    Exonerate for alignment anchoring

Correcting (arbitrarily) for multifurcating nodes.
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 1.

Alignment score: 9063
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 2.

Alignment score: 7787
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 3.

Alignment score: 8220
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 4.

Alignment score: 7867
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 5.

Alignment score: 7869


Writing
 - alignment to 'SUP35_250seqs_prank.fa.best.fas'

Analysis done. Total time 119s

116.98user 1.16system 1:59.50elapsed 98%CPU (0avgtext+0avgdata 66236maxresident)k
0inputs+16704outputs (0major+629152minor)pagefaults 0swaps
```

In [29]:
```
record_dict = IO.to_dict(IO.parse("SUP35_250seqs_prank.fa.best.fas", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
    break
```

```
SUP35_Agos_ATCC_10895_NM_211584
  2472
```

Compare algorithms on 250 sequences

algorithm	alignment length (nt)	time (seconds)
ClustalW	2179	1552.93
MUSCLE	2313	93.33
MAFFT	2322	15.11
Kalign	2214	15.07
t-coffee	-	-
prank	2472	102.53

Conclusion: here we can highlight two algorithms - **MAFFT** (it is the fastest one and has high alignment length) and **Prank** - it is a bit slower but has the highest length of alignment

Translation into amino acids

In [30]:
```
# sudo apt-get update -y
# sudo apt-get install -y emboss
```

Transeq translates nucleotide sequence into amino acids just directly reading codons

In [31]:
```
! transeq -sequence SUP35_10seqs.fa -outseq SUP35_10seqs.t.faa
```

```
Translate nucleic acid sequences
```

In [32]:
```
! head -20 SUP35_10seqs.t.faa
```

```
>SUP35_Kla_AB039749_1
MSDQQNQDQGQGQGYNQYNQYGQYNQYYNQQGYQGYNGQQGAPQGYQAYQAYGQQPQGAY
QGYNPQQAQGYQPYQGYNAQQQGYNAQQGGHNNNYNKNYNNKNSYNNYNKQGYQGAQGYN
AQQPTGYAAPAQSSSQGMTLKDFQNQQGSTNAAKPKPKLKLASSSGIKLVGAKKPVAPKT
EKTDESKEATKTTDDNEEAQSELPKIDDLKISEAEKPKTKENTPSADDTSSEKTTSAKAD
TSTGGANSVDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLTGS
VDKRTVEKYEREAKEAGRQGWYLSWVMDTNKEERNDGKTIEVGRAYFETEKRRYTILDAP
GHKMYVSEMIGGASQADIGILVISARKGEYETGFEKGGQTREHALLAKTQGVNKMIVVIN
KMDDPTVGWDKERYDHCVGNLTNFLKAVGYNVKEDVIFMPVSGYTGAGLKERVDPKDCPW
YTGPSLLEYLDNMKTTDRHINAPFMLPIASKMKDMGTVVEGKIESGHIRKGNQTLLMPNR
TSVEILTIYNETESEVDMAVCGEQVRLRIKGVEEEEISAGFVLTSPKNPVKNVTRFVAQI
AIVELKSIMSAGFSCVMHIHTAIEEVTVTRLLHKLEKGSNRKSKKPPAFAKKGMKIIAVI
ETNEPVCVETYDDYPQLGRFTLRDQGTTIAIGKIVKILEN*
>SUP35_Agos_ATCC_10895_NM_211584_1
MSEEDQIQSQGNDQGQSQAKDQGQNQGQGQQNFGQYYNPSNFQNYQGYVPQGGYQAYGQQ
AGGYQGYAQYNQQAGGYQGYQGYQQYNPAQAGYQGYQQYNAQGGYQSYKQYNSQPQGNRK
GNQSYGYGQQSATAPVTLNNFEKGTVPNATAPKPKKTLKLASSSGIKLVGAKKPVAKKE
EAKAEEPTKEEPKSSAEGAPKSEDATASSEDKAVPSIEKLSISEADTAKKDTADAAGATS
SDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLTGSVDKRTVEK
YEKEAKEAGRQGWYLSWIMDTNKEERNDGKTIEVGRSYFETEKRRYTILDAPGHKMYVSE
```

Getorf searches for ORFs in an input file and then outputs them.

`-minsize` : minimum size of ORF (otherwise it will output too many ORFs per sequence and most of them will be a garbage)

In [33]:
```
! getorf -sequence SUP35_10seqs.fa -outseq SUP35_10seqs.g.faa -noreverse -minsize 500
```

```
Find and extract open reading frames (ORFs)
```

[...] - coordinates of ORFs

In [34]:
```
! head -20 SUP35_10seqs.g.faa
```

```
>SUP35_Kla_AB039749_1 [1 - 2100]
MSDQQNQDQGQGQGYNQYNQYGQYNQYYNQQGYQGYNGQQGAPQGYQAYQAYGQQPQGAY
QGYNPQQAQGYQPYQGYNAQQQGYNAQQGGHNNNYNKNYNNKNSYNNYNKQGYQGAQGYN
AQQPTGYAAPAQSSSQGMTLKDFQNQQGSTNAAKPKPKLKLASSSGIKLVGAKKPVAPKT
EKTDESKEATKTTDDNEEAQSELPKIDDLKISEAEKPKTKENTPSADDTSSEKTTSAKAD
TSTGGANSVDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLTGS
VDKRTVEKYEREAKEAGRQGWYLSWVMDTNKEERNDGKTIEVGRAYFETEKRRYTILDAP
GHKMYVSEMIGGASQADIGILVISARKGEYETGFEKGGQTREHALLAKTQGVNKMIVVIN
KMDDPTVGWDKERYDHCVGNLTNFLKAVGYNVKEDVIFMPVSGYTGAGLKERVDPKDCPW
YTGPSLLEYLDNMKTTDRHINAPFMLPIASKMKDMGTVVEGKIESGHIRKGNQTLLMPNR
TSVEILTIYNETESEVDMAVCGEQVRLRIKGVEEEEISAGFVLTSPKNPVKNVTRFVAQI
AIVELKSIMSAGFSCVMHIHTAIEEVTVTRLLHKLEKGSNRKSKKPPAFAKKGMKIIAVI
ETNEPVCVETYDDYPQLGRFTLRDQGTTIAIGKIVKILEN
>SUP35_Agos_ATCC_10895_NM_211584_1 [1 - 2073]
MSEEDQIQSQGNDQGQSQAKDQGQNQGQGQQNFGQYYNPSNFQNYQGYVPQGGYQAYGQQ
AGGYQGYAQYNQQAGGYQGYQGYQQYNPAQAGYQGYQQYNAQGGYQSYKQYNSQPQGNRK
GNQSYGYGQQSATAPVTLNNFEKGTVPNATAPKPKKTLKLASSSGIKLVGAKKPVAKKE
EAKAEEPTKEEPKSSAEGAPKSEDATASSEDKAVPSIEKLSISEADTAKKDTADAAGATS
SDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLTGSVDKRTVEK
YEKEAKEAGRQGWYLSWIMDTNKEERNDGKTIEVGRSYFETEKRRYTILDAPGHKMYVSE
```

Protein alignments

ClustalW

`-TYPE` : specifies a type of the sequence.
Other algorithms can identify it automatically just reading the first letters of input sequences

In [35]:
```
! time clustalw -INFILE=SUP35_10seqs.g.faa -OUTFILE=SUP35_10seqs_clustalw.faa -OUTPUT=FASTA -TYPE=pro
tein
```

```


 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: SUP35_Kla_AB039749_1                      700 aa
Sequence 2: SUP35_Agos_ATCC_10895_NM_211584_1         691 aa
Sequence 3: SUP35_Scer_74-D694_GCA_001578265.1_1      685 aa
Sequence 4: SUP35_Spar_A12_Liti_1                     681 aa
Sequence 5: SUP35_Smik_IFO1815T_30_1                  681 aa
Sequence 6: SUP35_Skud_IFO1802T_36_1                  678 aa
Sequence 7: SUP35_Sbou_unique28_CM003560_1            685 aa
Sequence 8: SUP35_Scer_beer078_CM005938_1             666 aa
Sequence 9: SUP35_Sarb_H-6_chrXIII_CM001575_1         680 aa
Sequence 10: SUP35_Seub_CBS12357_chr_II_IV_DF968535_1   674 aa
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  71
Sequences (1:3) Aligned. Score:  69
Sequences (1:4) Aligned. Score:  69
Sequences (1:5) Aligned. Score:  68
Sequences (1:6) Aligned. Score:  68
Sequences (1:7) Aligned. Score:  69
Sequences (1:8) Aligned. Score:  70
Sequences (1:9) Aligned. Score:  67
Sequences (1:10) Aligned. Score:  69
Sequences (2:3) Aligned. Score:  68
Sequences (2:4) Aligned. Score:  69
Sequences (2:5) Aligned. Score:  70
Sequences (2:6) Aligned. Score:  69
Sequences (2:7) Aligned. Score:  69
Sequences (2:8) Aligned. Score:  71
Sequences (2:9) Aligned. Score:  67
Sequences (2:10) Aligned. Score:  69
Sequences (3:4) Aligned. Score:  97
Sequences (3:5) Aligned. Score:  94
Sequences (3:6) Aligned. Score:  92
Sequences (3:7) Aligned. Score:  99
Sequences (3:8) Aligned. Score:  99
Sequences (3:9) Aligned. Score:  92
Sequences (3:10) Aligned. Score:  91
Sequences (4:5) Aligned. Score:  94
Sequences (4:6) Aligned. Score:  92
Sequences (4:7) Aligned. Score:  97
Sequences (4:8) Aligned. Score:  96
Sequences (4:9) Aligned. Score:  93
Sequences (4:10) Aligned. Score:  91
Sequences (5:6) Aligned. Score:  92
Sequences (5:7) Aligned. Score:  94
Sequences (5:8) Aligned. Score:  94
Sequences (5:9) Aligned. Score:  92
Sequences (5:10) Aligned. Score:  90
Sequences (6:7) Aligned. Score:  92
Sequences (6:8) Aligned. Score:  91
Sequences (6:9) Aligned. Score:  91
Sequences (6:10) Aligned. Score:  90
Sequences (7:8) Aligned. Score:  100
Sequences (7:9) Aligned. Score:  93
Sequences (7:10) Aligned. Score:  91
Sequences (8:9) Aligned. Score:  92
Sequences (8:10) Aligned. Score:  91
Sequences (9:10) Aligned. Score:  91
Guide tree file created:   [SUP35_10seqs.g.dnd]

There are 9 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:14421
Group 2: Sequences:   3      Score:14556
Group 3: Sequences:   4      Score:14367
Group 4: Sequences:   5      Score:14254
Group 5: Sequences:   6      Score:14159
Group 6: Sequences:   7      Score:14064
Group 7: Sequences:   8      Score:13830
Group 8: Sequences:   2      Score:12740
Group 9: Sequences:  10      Score:11985
Alignment Score 155481
firstres = 1 lastres = 719
FASTA file created!

Fasta-Alignment file created    [SUP35_10seqs_clustalw.faa]

0.39user 0.00system 0:00.39elapsed 100%CPU (0avgtext+0avgdata 4852maxresident)k
0inputs+24outputs (0major+298minor)pagefaults 0swaps
```

In [36]:
```
! head -20 SUP35_10seqs_clustalw.faa
```

```
>SUP35_Scer_74-D694_GCA_001578265.1_1
MSDSN----QGNNQQNYQQYSQNGNQQQG----NNRY--Q
GYQAYNAQAQ-PAGGYYQNYQGYSGYQQGGY----QQYNP
DAGYQQQYNPQGGYQQYNP-QGGYQQ----QFNPQGGRGN
YKNFNYNNNLQ-------GYQAGFQPQSQGMSLNDFQK--
QQKQAAPKPKKTLKLVSSSGIKLANATKKVGTKPAESDKK
EEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEEKTEEKSE
LPKVEDLKISEST----HNTNNANVTSADALIKEQEEEVD
DEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLTGSV
DKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIE
VGKAYFETEKRRYTILDAPGHKMYVSEMIGGASQADVGVL
VISARKGEYETGFERGGQTREHALLAKTQGVNKMVVVVNK
MDDPTVNWSKERYDQCVSNVSNFLRAIGYNIKTDVVFMPV
SGYSGANLKDHVDPKECPWYTGPTLLEYLDTMNHVDRHIN
APFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKT
AVEIQNIYNETENEVDMAMCGEQVKLRIKGVEEEDISPGF
VLTSPKNPIKSVTKFVAQIAIVELKSIIAAGFSCVMHVHT
AIEEVHIVKLLHKLEKGTNRKSKKPPAFAKKGMKVIAVLE
TEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIAE-
>SUP35_Scer_beer078_CM005938_1
```

In [37]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_clustalw.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Scer_74-D694_GCA_001578265.1_1
  719
SUP35_Scer_beer078_CM005938_1
  719
SUP35_Sbou_unique28_CM003560_1
  719
SUP35_Spar_A12_Liti_1
  719
SUP35_Smik_IFO1815T_30_1
  719
SUP35_Sarb_H-6_chrXIII_CM001575_1
  719
SUP35_Skud_IFO1802T_36_1
  719
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  719
SUP35_Kla_AB039749_1
  719
SUP35_Agos_ATCC_10895_NM_211584_1
  719
```

Clustal Omega

In [38]:
```
! time clustalo --infile=SUP35_10seqs.g.faa --outfile=SUP35_10seqs_clustalo.faa --verbose
```

```
Using 8 threads
Read 10 sequences (type: Protein) from SUP35_10seqs.g.faa
not more sequences (10) than cluster-size (100), turn off mBed
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress done. CPU time: 0.11u 0.00s 00:00:00.11 Elapsed: 00:00:00
Guide-tree computation done.
Progressive alignment progress done. CPU time: 1.05u 0.03s 00:00:01.08 Elapsed: 00:00:00
Alignment written to SUP35_10seqs_clustalo.faa
1.16user 0.03system 0:00.48elapsed 251%CPU (0avgtext+0avgdata 32988maxresident)k
1784inputs+16outputs (11major+44463minor)pagefaults 0swaps
```

In [39]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_clustalo.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1
  757
SUP35_Agos_ATCC_10895_NM_211584_1
  757
SUP35_Scer_74-D694_GCA_001578265.1_1
  757
SUP35_Spar_A12_Liti_1
  757
SUP35_Smik_IFO1815T_30_1
  757
SUP35_Skud_IFO1802T_36_1
  757
SUP35_Sbou_unique28_CM003560_1
  757
SUP35_Scer_beer078_CM005938_1
  757
SUP35_Sarb_H-6_chrXIII_CM001575_1
  757
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  757
```

Muscle

In [40]:
```
! time muscle -in SUP35_10seqs.g.faa -out SUP35_10seqs_muscle.faa
```

```
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

SUP35_10seqs.g 10 seqs, lengths min 666, max 700, avg 682
00:00:00    15 MB(-2%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    15 MB(-2%)  Iter   1  100.00%  K-mer dist pass 2
00:00:01    22 MB(-3%)  Iter   1  100.00%  Align node
00:00:01    22 MB(-3%)  Iter   1  100.00%  Root alignment
00:00:01    22 MB(-3%)  Iter   2  100.00%  Refine tree
00:00:01    22 MB(-3%)  Iter   2  100.00%  Root alignment
00:00:01    22 MB(-3%)  Iter   2  100.00%  Root alignment
00:00:01    22 MB(-3%)  Iter   3  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   4  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   5  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   5  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   6  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   7  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   8  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   8  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter   9  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  10  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  11  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  12  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  13  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  13  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  14  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  15  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  16  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  17  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  18  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  19  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  20  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  21  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  22  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  23  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  24  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  25  100.00%  Refine biparts
00:00:01    22 MB(-3%)  Iter  26  100.00%  Refine biparts
0.15user 0.00system 0:00.15elapsed 99%CPU (0avgtext+0avgdata 10784maxresident)k
0inputs+16outputs (0major+2094minor)pagefaults 0swaps
```

In [41]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_muscle.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1
  743
SUP35_Agos_ATCC_10895_NM_211584_1
  743
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  743
SUP35_Skud_IFO1802T_36_1
  743
SUP35_Sarb_H-6_chrXIII_CM001575_1
  743
SUP35_Smik_IFO1815T_30_1
  743
SUP35_Spar_A12_Liti_1
  743
SUP35_Scer_74-D694_GCA_001578265.1_1
  743
SUP35_Sbou_unique28_CM003560_1
  743
SUP35_Scer_beer078_CM005938_1
  743
```

MAFFT

In [42]:
```
! time mafft --auto SUP35_10seqs.g.faa >SUP35_10seqs_mafft.faa
```

```
outputhat23=16
treein = 0
compacttree = 0
stacksize: 8192 kb
rescale = 1
All-to-all alignment.
tbfast-pair (aa) Version 7.453
alg=L, model=BLOSUM62, 2.00, -0.10, +0.10, noshift, amax=0.0
0 thread(s)

outputhat23=16
Loading 'hat3.seed' ...
done.
Writing hat3 for iterative refinement
rescale = 1
Gap Penalty = -1.53, +0.00, +0.00
tbutree = 1, compacttree = 0
Constructing a UPGMA tree ...
    0 / 10
done.

Progressive alignment ...
STEP     8 /9
Reallocating..done. *alloclen = 2413
STEP     9 /9
done.
tbfast (aa) Version 7.453
alg=A, model=BLOSUM62, 1.53, -0.00, -0.00, noshift, amax=0.0
1 thread(s)

minimumweight = 0.000010
autosubalignment = 0.000000
nthread = 0
randomseed = 0
blosum 62 / kimura 200
poffset = 0
niter = 16
sueff_global = 0.100000
nadd = 16
Loading 'hat3' ... done.
rescale = 1

    0 / 10
Segment   1/  1    1- 753
STEP 003-003-1  identical.    accepted.
Converged.

done
dvtditr (aa) Version 7.453
alg=A, model=BLOSUM62, 1.53, -0.00, -0.00, noshift, amax=0.0
0 thread(s)


Strategy:
 L-INS-i (Probably most accurate, very slow)
 Iterative refinement method (<16) with LOCAL pairwise alignment information

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.

0.41user 0.03system 0:00.40elapsed 111%CPU (0avgtext+0avgdata 14484maxresident)k
1256inputs+152outputs (9major+30851minor)pagefaults 0swaps
```

In [43]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_mafft.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1
  759
SUP35_Agos_ATCC_10895_NM_211584_1
  759
SUP35_Scer_74-D694_GCA_001578265.1_1
  759
SUP35_Spar_A12_Liti_1
  759
SUP35_Smik_IFO1815T_30_1
  759
SUP35_Skud_IFO1802T_36_1
  759
SUP35_Sbou_unique28_CM003560_1
  759
SUP35_Scer_beer078_CM005938_1
  759
SUP35_Sarb_H-6_chrXIII_CM001575_1
  759
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  759
```

Kalign

In [44]:
```
! time kalign <SUP35_10seqs.g.faa >SUP35_10seqs_kalign.faa
```

```
Kalign version 2.04, Copyright (C) 2004, 2005, 2006 Timo Lassmann

        Kalign is free software. You can redistribute it and/or modify
        it under the terms of the GNU General Public License as
        published by the Free Software Foundation.

reading from STDIN: found 10 sequences
Aligning 10 protein sequences with these parameters:
        54.94940948     gap open penalty
        8.52492046      gap extension
        4.42409992      terminal gap penalty
        0.20000000      bonus
Alignment will be written to stdout.

Distance Calculation:
     100 percent done
Alignment:
     100 percent done
0.03user 0.00system 0:00.04elapsed 90%CPU (0avgtext+0avgdata 2888maxresident)k
0inputs+16outputs (0major+594minor)pagefaults 0swaps
```

In [45]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_kalign.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1
  720
SUP35_Agos_ATCC_10895_NM_211584_1
  720
SUP35_Scer_74-D694_GCA_001578265.1_1
  720
SUP35_Spar_A12_Liti_1
  720
SUP35_Smik_IFO1815T_30_1
  720
SUP35_Skud_IFO1802T_36_1
  720
SUP35_Sbou_unique28_CM003560_1
  720
SUP35_Scer_beer078_CM005938_1
  720
SUP35_Sarb_H-6_chrXIII_CM001575_1
  720
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  720
```

t-coffee

In [46]:
```
! time t_coffee -infile=SUP35_10seqs.g.faa -outfile=SUP35_10seqs_tcoffee.faa -output=fasta_aln
```

```
*****************************************************************************
Command line arguments: ['/home/isemenov/anaconda3/bin/t_coffee', '-infile=SUP35_10seqs.g.faa', '-out
file=SUP35_10seqs_tcoffee.faa', '-output=fasta_aln']
Install folder: /home/isemenov/anaconda3/lib/t_coffee-11.0.8
*****************************************************************************


PROGRAM: T-COFFEE Version_11.00.8cbe486 (2014-08-12 22:05:29 - Revision 8cbe486 - Build 477)
-full_log       S       [0]
-genepred_score S       [0]     nsd
-run_name       S       [0]
-mem_mode       S       [0]     mem
-extend         D       [1]     1
-extend_mode    S       [0]     very_fast_triplet
-max_n_pair     D       [0]     10
-seq_name_for_quadruplet        S       [0]     all
-compact        S       [0]     default
-clean          S       [0]     no
-do_self        FL      [0]     0
-do_normalise   D       [0]     1000
-template_file  S       [0]
-setenv         S       [0]     0
-template_mode  S       [0]
-flip           D       [0]     0
-remove_template_file   D       [0]     0
-profile_template_file  S       [0]
-in             S       [0]
-seq            S       [0]
-aln            S       [0]
-method_limits  S       [0]
-method         S       [0]
-lib            S       [0]
-profile        S       [0]
-profile1       S       [0]
-profile2       S       [0]
-pdb            S       [0]
-relax_lib      D       [0]     1
-filter_lib     D       [0]     0
-shrink_lib     D       [0]     0
-out_lib        W_F     [0]     no
-out_lib_mode   S       [0]     primary
-lib_only       D       [0]     0
-outseqweight   W_F     [0]     no
-dpa            FL      [0]     0
-seq_source     S       [0]     ANY
-cosmetic_penalty       D       [0]     0
-gapopen        D       [0]     0
-gapext         D       [0]     0
-fgapopen       D       [0]     0
-fgapext        D       [0]     0
-nomatch        D       [0]     0
-newtree        W_F     [0]     default
-tree           W_F     [0]     NO
-usetree        R_F     [0]
-tree_mode      S       [0]     nj
-distance_matrix_mode   S       [0]     ktup
-distance_matrix_sim_mode       S       [0]     idmat_sim1
-quicktree      FL      [0]     0
-outfile        W_F     [1]     SUP35_10seqs_tcoffee.faa
-maximise       FL      [1]     1
-output         S       [1]     fasta_aln
-len            D       [0]     0
-infile         R_F     [1]     SUP35_10seqs.g.faa
-matrix         S       [0]     default
-tg_mode        D       [0]     1
-profile_mode   S       [0]     cw_profile_profile
-profile_comparison     S       [0]     profile
-dp_mode        S       [0]     linked_pair_wise
-ktuple         D       [0]     1
-ndiag          D       [0]     0
-diag_threshold D       [0]     0
-diag_mode      D       [0]     0
-sim_matrix     S       [0]     vasiliky
-transform      S       [0]
-extend_seq     FL      [0]     0
-outorder       S       [0]     input
-inorder        S       [0]     aligned
-seqnos         S       [0]     off
-case           S       [0]     keep
-cpu            D       [0]     0
-ulimit         D       [0]     -1
-maxnseq        D       [0]     -1
-maxlen         D       [0]     -1
-sample_dp      D       [0]     0
-weight         S       [0]     default
-seq_weight     S       [0]     no
-align          FL      [1]     1
-mocca          FL      [0]     0
-domain         FL      [0]     0
-start          D       [0]     0
-len            D       [0]     0
-scale          D       [0]     0
-mocca_interactive      FL      [0]     0
-method_evaluate_mode   S       [0]     default
-evaluate_mode  S       [0]     triplet
-get_type       FL      [0]     0
-clean_aln      D       [0]     0
-clean_threshold        D       [1]     1
-clean_iteration        D       [1]     1
-clean_evaluate_mode    S       [0]     t_coffee_fast
-extend_matrix  FL      [0]     0
-prot_min_sim   D       [40]    40
-prot_max_sim   D       [90]    90
-prot_min_cov   D       [40]    40
-pdb_type       S       [0]     d
-pdb_min_sim    D       [35]    35
-pdb_max_sim    D       [100]   100
-pdb_min_cov    D       [50]    50
-pdb_blast_server       W_F     [0]     EBI
-blast          W_F     [0]
-blast_server   W_F     [0]     EBI
-pdb_db         W_F     [0]     pdb
-protein_db     W_F     [0]     uniprot
-method_log     W_F     [0]     no
-struc_to_use   S       [0]
-cache          W_F     [0]     use
-align_pdb_param_file   W_F     [0]     no
-align_pdb_hasch_mode   W_F     [0]     hasch_ca_trace_bubble
-external_aligner       S       [0]     NO
-msa_mode       S       [0]     tree
-master         S       [0]     no
-blast_nseq     D       [0]     0
-lalign_n_top   D       [0]     10
-iterate        D       [0]     0
-trim           D       [0]     0
-split          D       [0]     0
-trimfile       S       [0]     default
-split          D       [0]     0
-split_nseq_thres       D       [0]     0
-split_score_thres      D       [0]     0
-check_pdb_status       D       [0]     0
-clean_seq_name D       [0]     0
-seq_to_keep    S       [0]
-dpa_master_aln S       [0]
-dpa_maxnseq    D       [0]     0
-dpa_min_score1 D       [0]
-dpa_min_score2 D       [0]
-dpa_keep_tmpfile       FL      [0]     0
-dpa_debug      D       [0]     0
-multi_core     S       [0]     templates_jobs_relax_msa_evaluate
-n_core         D       [0]     0
-max_n_proc     D       [0]     0
-lib_list       S       [0]
-prune_lib_mode S       [0]     5
-tip            S       [0]     none
-rna_lib        S       [0]
-no_warning     D       [0]     0
-run_local_script       D       [0]     0
-plugins        S       [0]     default
-proxy          S       [0]     unset
-email          S       [0]
-clean_overaln  D       [0]     0
-overaln_param  S       [0]
-overaln_mode   S       [0]
-overaln_model  S       [0]
-overaln_threshold      D       [0]     0
-overaln_target D       [0]     0
-overaln_P1     D       [0]     0
-overaln_P2     D       [0]     0
-overaln_P3     D       [0]     0
-overaln_P4     D       [0]     0
-exon_boundaries        S       [0]
-dump           S       [0]     no
-display        D       [0]     100

INPUT FILES
        Input File (S) SUP35_10seqs.g.faa  Format fasta_seq
        Input File (M) proba_pair

Identify Master Sequences [no]:

Master Sequences Identified
INPUT SEQUENCES: 10 SEQUENCES  [PROTEIN]
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Agos_ATCC_10895_NM_211584_1        Le
ngth  691 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Kla_AB039749_1                     Le
ngth  700 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Sarb_H-6_chrXIII_CM001575_1        Le
ngth  680 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Sbou_unique28_CM003560_1           Le
ngth  685 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Scer_74-D694_GCA_001578265.1_1     Le
ngth  685 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Scer_beer078_CM005938_1            Le
ngth  666 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Seub_CBS12357_chr_II_IV_DF968535_1 Le
ngth  674 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Skud_IFO1802T_36_1                 Le
ngth  678 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Smik_IFO1815T_30_1                 Le
ngth  681 type PROTEIN Struct Unchecked
  Input File SUP35_10seqs.g.faa                         Seq SUP35_Spar_A12_Liti_1                    Le
ngth  681 type PROTEIN Struct Unchecked

        Multi Core Mode: 8 processors:

        --- Process Method/Library/Aln SSUP35_10seqs.g.faa
        --- Process Method/Library/Aln Mproba_pair
        xxx Retrieved SSUP35_10seqs.g.faa
        xxx Retrieved Mproba_pair

        All Methods Retrieved

MANUAL PENALTIES: gapopen=0 gapext=0

        Library Total Size: [87364]

Library Relaxation: Multi_proc [8]

!               [Relax Library][TOT=    1][100 %][ ELAPSED  TIME:    0 sec.]
Relaxation Summary: [87364]--->[77444]



UN-WEIGHTED MODE: EVERY SEQUENCE WEIGHTS 1

MAKE GUIDE TREE
        [MODE=nj][DONE]

PROGRESSIVE_ALIGNMENT [Tree Based]
Group    1: SUP35_Agos_ATCC_10895_NM_211584_1
Group    2: SUP35_Kla_AB039749_1
Group    3: SUP35_Sarb_H-6_chrXIII_CM001575_1
Group    4: SUP35_Sbou_unique28_CM003560_1
Group    5: SUP35_Scer_74-D694_GCA_001578265.1_1
Group    6: SUP35_Scer_beer078_CM005938_1
Group    7: SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
Group    8: SUP35_Skud_IFO1802T_36_1
Group    9: SUP35_Smik_IFO1815T_30_1
Group   10: SUP35_Spar_A12_Liti_1


        Group   11: [Group    5 (   1 seq)] with [Group    4 (   1 seq)]-->[Len=  685][PID:28734]
        Group   12: [Group    6 (   1 seq)] with [Group   11 (   2 seq)]-->[Len=  685][PID:28734]
        Group   13: [Group   12 (   3 seq)] with [Group   10 (   1 seq)]-->[Len=  685][PID:28734]
        Group   14: [Group    9 (   1 seq)] with [Group   13 (   4 seq)]-->[Len=  685][PID:28734]
        Group   15: [Group    8 (   1 seq)] with [Group   14 (   5 seq)]-->[Len=  685][PID:28734]
        Group   16: [Group    3 (   1 seq)] with [Group   15 (   6 seq)]-->[Len=  685][PID:28734]
        Group   17: [Group    7 (   1 seq)] with [Group   16 (   7 seq)]-->[Len=  687][PID:28734]
        Group   19: [Group   18 (   2 seq)] with [Group   17 (   8 seq)]-->[Len=  752][PID:28709][For
ked]


!               [Final Evaluation][TOT=   94][100 %][ ELAPSED  TIME:    0 sec.]
CLUSTAL FORMAT for T-COFFEE Version_11.00.8cbe486 [http://www.tcoffee.org] [MODE:  ], CPU=0.00 sec, S
CORE=958, Nseq=10, Len=752

SUP35_Kla_AB039749_1                        MSDQQNQDQGQGQGYNQYNQYGQYNQYYNQ-QGYQGYNGQQ-GAP-QGYQ
SUP35_Agos_ATCC_10895_NM_211584_1           MSEED-QIQSQ-GNDQ--GQSQAKDQGQNQGQGQQNFGQYY-NPS--NFQ
SUP35_Scer_74-D694_GCA_001578265.1_1        MSDSN-QGNNQ-QNYQQYSQNGNQQQGNNRYQGYQAYNAQA-QPAGGYYQ
SUP35_Spar_A12_Liti_1                       MSDSN-QGNNQ-QSYQQYGQNPNQQQGNNRYQGYQAYNAQS-QPAGGYYQ
SUP35_Smik_IFO1815T_30_1                    MSDSN-QGNNQ-QNYQQYNQNPNQQQGNNRYQGYQAYNAQA-QPASGHYQ
SUP35_Skud_IFO1802T_36_1                    MSDPN-QGNNQ-Q---QYGQNPNQQQGNNKYQGYQAYNAQA-QPANGYYQ
SUP35_Sbou_unique28_CM003560_1              MSDSN-QGNNQ-QNYQQYSQNGNQQQGNNRYQGYQAYNAQA-QPAGGYYQ
SUP35_Scer_beer078_CM005938_1               MSDSN-QGNNQ-QNYQQYSQNGNQQQGNNRYQGYQAYNAQA-QPAGGYYQ
SUP35_Sarb_H-6_chrXIII_CM001575_1           MSDPT-NGNNE-QSSQQQGQNPSQQQGNNRYQGYQAYNAQT-QPAGGYYQ
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    MSDPN-QGNNQ-QNYQQYGQNFNQQQGNNKFQGYQAYNAQAQQPAGGYYQ
                                            **:   : :.:       .*    :*  *: ** * :.     ..   :*

SUP35_Kla_AB039749_1                        AYQAYGQQPQGAY-QGYNPQQA--Q---GYQPY-------QGY-------
SUP35_Agos_ATCC_10895_NM_211584_1           NYQGYV--PQGGY-QAYGQQAGGYQGYAQYNQQAGGYQGYQGYQQYNPAQ
SUP35_Scer_74-D694_GCA_001578265.1_1        NYQGYSGYQQGGY-QQYNPDAGYQQ---QYNPQ-------GGYQQYNP-Q
SUP35_Spar_A12_Liti_1                       NYQGYSGYQQGSY-QQYNPQQGYQQ---QYNPQ-------GGYQQYNP-Q
SUP35_Smik_IFO1815T_30_1                    NYQGYSGYQQGAF-QQYNPEAGYQQ---QYNPQ-------GGYQQYNS-Q
SUP35_Skud_IFO1802T_36_1                    NYQGFAGYQQGGY-QQYDRDAGSQQ---QYNPQ-------GGYQQYPP-Q
SUP35_Sbou_unique28_CM003560_1              NYQGYSGYQQGGY-QQYNPDAGYQQ---QYNPQ-------GGYQQYNP-Q
SUP35_Scer_beer078_CM005938_1               NYQGYSGYQQGGY-Q---------------------------QYNP-Q
SUP35_Sarb_H-6_chrXIII_CM001575_1           NYQGYAGYQQGGY-QQFTPEAGYQQ---QYNPQ-------GGYQQFNP-Q
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    NPQGYAGYQQGGYDQQFNPEVGYQQ------------------QYNA-Q
                                              *.:    **.: *

SUP35_Kla_AB039749_1                        NAQQ--QGYNAQQGGHNNNYNKNYNNKNSYNNYNKQGYQGAQGYNAQQPT
SUP35_Agos_ATCC_10895_NM_211584_1           AGYQGYQQYNA-QGGYQSYKQYN--S-------QPQGNRK--GN---QSY
SUP35_Scer_74-D694_GCA_001578265.1_1        GGYQ--QQFNP-QGGRGNYKNFNYNN-------NL--------------Q
SUP35_Spar_A12_Liti_1                       GGYQ--QQFNP-QGGRGNYKNFNYNN-------NA--------------Q
SUP35_Smik_IFO1815T_30_1                    GGYQ--QQFNP-QGGRGNYQNINYNS-------NI--------------Q
SUP35_Skud_IFO1802T_36_1                    EGYQ--QQSNP-QGGRGNYKSFNYNN-------NV--------------Q
SUP35_Sbou_unique28_CM003560_1              GGYQ--QQFNP-QGGRGNYKNFNYNN-------NL--------------Q
SUP35_Scer_beer078_CM005938_1               GGYQ--QQFNP-QGGRGNYKNFNYNN-------NL--------------Q
SUP35_Sarb_H-6_chrXIII_CM001575_1           GGYQ--QQFNP-QGGRGNYKNFN-NN-------NQ--------------Q
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    GGYQ--QQFNP-QGGRGNYKNFNYNN-------SQ--------------Q
                                             . *  *  *. ***  .  . *  .        .

SUP35_Kla_AB039749_1                        GYAAPAQSSSQGMTLKDFQNQQG-STNAAKPKPKLKLASSSGIKLVGAKK
SUP35_Agos_ATCC_10895_NM_211584_1           GYGQG-QSATAPVTLNNFEKGTVPNATAPKPKKTLKLASSSGIKLVGAKK
SUP35_Scer_74-D694_GCA_001578265.1_1        GYQAGFQPQSQGMSLNDFQKQQK--QAAPKPKKTLKLVSSSGIKLANATK
SUP35_Spar_A12_Liti_1                       GYQAGFQPQSQGMSLNDFQKQQK--QAAPKPKKTLKLVSSSGIKLANATK
SUP35_Smik_IFO1815T_30_1                    GYQAGFQPQSQSMSLNDFQKQQK--QAAPKPKKTLKLVSSSGIKLANATK
SUP35_Skud_IFO1802T_36_1                    GYQAGFQPQSQGMSLNDFQKQQK--QTAPKPKKTLKLVSSSGIKLANATK
SUP35_Sbou_unique28_CM003560_1              GYQAGFQPQSQGMSLNDFQKQQK--QAAPKPKKTLKLVSSSGIKLANATK
SUP35_Scer_beer078_CM005938_1               GYQAGFQPQSQGMSLNDFQKQQK--QAAPKPKKTLKLVSSSGIKLANATK
SUP35_Sarb_H-6_chrXIII_CM001575_1           GYQSNFQPQSQGMSLNDFQKQQK--QSTPKPKKTLKLVSSSGIKLANATK
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    GFQAGFQPQSQGMSLNDFQKQQK--QTAPKPKKTLKLVSSSGIKLANATK
                                            *:    *. :  ::*::*::       :.*** .***.*******..*.*

SUP35_Kla_AB039749_1                        PVAPKTEKTDE--------SKE----ATK-----------TTDDN--EE
SUP35_Agos_ATCC_10895_NM_211584_1           PVAKKEEA--K--------AEE----PTK-EEPKSSAEGAPKSEDATASS
SUP35_Scer_74-D694_GCA_001578265.1_1        KVGTKPAESDKKEEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEEK--TE
SUP35_Spar_A12_Liti_1                       KVDTKPAESEKKKEDKPTENKE----PTKLEEPVKKEEKLVKTEEK--KE
SUP35_Smik_IFO1815T_30_1                    KVDTKPSETEKKEEVKPVETKK----PIKVEEPVKKEEMPVKAEEK--KE
SUP35_Skud_IFO1802T_36_1                    KVSTKPAETEKKEEDKPVKTEE----QIKVEEPIKEEKAVKTEEK--KE
SUP35_Sbou_unique28_CM003560_1              KVDTKPAESDKKEEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEEK--TE
SUP35_Scer_beer078_CM005938_1               KVDTKPAESDKKEEEKSAETKEPTKEPTKVEEPVKKEEKPVQTEEK--TE
SUP35_Sarb_H-6_chrXIII_CM001575_1           KVDTKPVETEKKEEDKPVETKK----PTKVEEPAKKEEKPVKAEEA--KE
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    KVDTKPVETEKKEEKPTETKE----PAKVEESIKDVETPASAEEK--KE
                                             *  *     :          ::      *              :::    .

SUP35_Kla_AB039749_1                        AQSELPKIDDLKISEAEKPKTKENTPSADDTSSEKTTSAKADTSTGGANS
SUP35_Agos_ATCC_10895_NM_211584_1           EDKAVPSIEKLSISEADTA------------------KKDTADAAGATS
SUP35_Scer_74-D694_GCA_001578265.1_1        EKSELPKVEDLKISESTH----------------------NTNNANVTS
SUP35_Spar_A12_Liti_1                       EKSELPKVEDLKISESTD----------------------NTNNANVNS
SUP35_Smik_IFO1815T_30_1                    EKSELPKVGDLKISESTD----------------------NSNTTSATS
SUP35_Skud_IFO1802T_36_1                    VESELPRVEDLKISESTD----------------------NSNTANVTS
SUP35_Sbou_unique28_CM003560_1              EKSELPKVEDLKISESTH----------------------NTNNANVTS
SUP35_Scer_beer078_CM005938_1               EKSELPKVEDLKISESTH----------------------NTNNANVTS
SUP35_Sarb_H-6_chrXIII_CM001575_1           EKSELPKVEDLKISEPTD----------------------NSNAASVNN
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    VNSELPKVEDLKISESND----------------------NSKPANPIN
                                             .. :* : .*.***.                          :.  .  .

SUP35_Kla_AB039749_1                        VDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLT
SUP35_Agos_ATCC_10895_NM_211584_1           SDALIKEQEDEVDEEVVKDMFGGKDHVSIIFMGHVDAGKSTMGGNLLYLT
SUP35_Scer_74-D694_GCA_001578265.1_1        ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Spar_A12_Liti_1                       ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Smik_IFO1815T_30_1                    ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Skud_IFO1802T_36_1                    ADTLIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Sbou_unique28_CM003560_1              ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Scer_beer078_CM005938_1               ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Sarb_H-6_chrXIII_CM001575_1           ADALIKEQEEEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    ADALIKEQEDEVDDEVVNDMFGGKDHVSLIFMGHVDAGKSTMGGNLLYLT
                                             *:******:***:***:**********:*********************

SUP35_Kla_AB039749_1                        GSVDKRTVEKYEREAKEAGRQGWYLSWVMDTNKEERNDGKTIEVGRAYFE
SUP35_Agos_ATCC_10895_NM_211584_1           GSVDKRTVEKYEKEAKEAGRQGWYLSWIMDTNKEERNDGKTIEVGRSYFE
SUP35_Scer_74-D694_GCA_001578265.1_1        GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Spar_A12_Liti_1                       GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Smik_IFO1815T_30_1                    GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Skud_IFO1802T_36_1                    GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Sbou_unique28_CM003560_1              GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Scer_beer078_CM005938_1               GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Sarb_H-6_chrXIII_CM001575_1           GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    GSVDKRTIEKYEREAKDAGRQGWYLSWVMDTNKEERNDGKTIEVGKAYFE
                                            *******:****:***:**********:*****************::***

SUP35_Kla_AB039749_1                        TEKRRYTILDAPGHKMYVSEMIGGASQADIGILVISARKGEYETGFEKGG
SUP35_Agos_ATCC_10895_NM_211584_1           TEKRRYTILDAPGHKMYVSEMIGGASQADVGILVISARKGEYETGFERGG
SUP35_Scer_74-D694_GCA_001578265.1_1        TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Spar_A12_Liti_1                       TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Smik_IFO1815T_30_1                    TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Skud_IFO1802T_36_1                    TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Sbou_unique28_CM003560_1              TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Scer_beer078_CM005938_1               TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Sarb_H-6_chrXIII_CM001575_1           TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    TEKRRYTILDAPGHKMYVSEMIGGASQADVGVLVISARKGEYETGFERGG
                                            *****************************:*:***************:**

SUP35_Kla_AB039749_1                        QTREHALLAKTQGVNKMIVVINKMDDPTVGWDKERYDHCVGNLTNFLKAV
SUP35_Agos_ATCC_10895_NM_211584_1           QTREHALLAKTQGVNKMVVVVNKMDDPTVNWDKARYDQCIKNVSNFLQAI
SUP35_Scer_74-D694_GCA_001578265.1_1        QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLRAI
SUP35_Spar_A12_Liti_1                       QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLRAI
SUP35_Smik_IFO1815T_30_1                    QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLKAI
SUP35_Skud_IFO1802T_36_1                    QTREHALLAKTQGVNKMVVVVNKMDDATVNWSKERYDQCVGNVSNFLKAI
SUP35_Sbou_unique28_CM003560_1              QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLRAI
SUP35_Scer_beer078_CM005938_1               QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLRAI
SUP35_Sarb_H-6_chrXIII_CM001575_1           QTREHALLAKTQGVNKMVVVVNKMDDPTVNWSKERYDQCVSNVSNFLKAI
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    QTREHALLAKTQGVNKMIVVVNKMDDPTVKWSKERYDQCVGNVSNFLKAI
                                            *****************:**:*****.** *.* ***:*: *::***:*:

SUP35_Kla_AB039749_1                        GYNVKEDVIFMPVSGYTGAGLKERVDPKDCPWYTGPSLLEYLDNMKTTDR
SUP35_Agos_ATCC_10895_NM_211584_1           GYNVKEDVMYMPVSGFTGAGLKDRVDKKDCPWYDGPSLLEYLDNMQHVDR
SUP35_Scer_74-D694_GCA_001578265.1_1        GYNIKTDVVFMPVSGYSGANLKDHVDPKECPWYTGPTLLEYLDTMNHVDR
SUP35_Spar_A12_Liti_1                       GYNIKTDVVFMPVSGYSGANLKDHVDPKECPWYTGPTLLEYLDTMNHVDR
SUP35_Smik_IFO1815T_30_1                    GYNIKTDVVFMPVSGYSGANLKNRVDPKECPWYTGPTLLEYLDTMNHVDR
SUP35_Skud_IFO1802T_36_1                    GYNIKTDVAFMPVSGYSGANLKNRVDPKECSWYTGPTLLEYLDTMNHVDR
SUP35_Sbou_unique28_CM003560_1              GYNIKTDVVFMPVSGYSGANLKDHVDPKECPWYTGPTLLEYLDTMNHVDR
SUP35_Scer_beer078_CM005938_1               GYNIKTDVVFMPVSGYSGANLKDHVDPKECPWYTGPTLLEYLDTMNHVDR
SUP35_Sarb_H-6_chrXIII_CM001575_1           GYNIKTDVVFMPVSGYSGANLKEHVNPKECSWYTGPTLLEYLDKMNHVDR
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    GYNIKTDVIFMPVSGYSGANLKEHVDPKECPWYTGPTLLEYLDKMTHVDR
                                            ***:* ** :*****::**.**::*: *:* ** **:******.*   .**

SUP35_Kla_AB039749_1                        HINAPFMLPIASKMKDMGTVVEGKIESGHIRKGNQTLLMPNRTSVEILTI
SUP35_Agos_ATCC_10895_NM_211584_1           FINAPFMLPIASKMKDMGTVVEGKIESGHIKKGNQTLLMPNKIPVEILAI
SUP35_Scer_74-D694_GCA_001578265.1_1        HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTAVEIQNI
SUP35_Spar_A12_Liti_1                       HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTAVEIQNI
SUP35_Smik_IFO1815T_30_1                    HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTAVEIQNI
SUP35_Skud_IFO1802T_36_1                    HVNAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTSVEIQNI
SUP35_Sbou_unique28_CM003560_1              HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTAVEIQNI
SUP35_Scer_beer078_CM005938_1               HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTAVEIQNI
SUP35_Sarb_H-6_chrXIII_CM001575_1           HINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTSVEIQNI
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    RINAPFMLPIAAKMKDLGTIVEGKIESGHIKKGQSTLLMPNKTTVEIQNI
                                             :*********:****:**:**********:**:.******: .***  *

SUP35_Kla_AB039749_1                        YNETESEVDMAVCGEQVRLRIKGVEEEEISAGFVLTSPKNPVKNVTRFVA
SUP35_Agos_ATCC_10895_NM_211584_1           QNETEQEVDMAVCGEQVRLRLKGVEEEDISAGFVLTSPKNPVKNVTKFVA
SUP35_Scer_74-D694_GCA_001578265.1_1        YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Spar_A12_Liti_1                       YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Smik_IFO1815T_30_1                    YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Skud_IFO1802T_36_1                    YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Sbou_unique28_CM003560_1              YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Scer_beer078_CM005938_1               YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Sarb_H-6_chrXIII_CM001575_1           YNETENEVDMAMCGEQVKLRIKGVEEEDISPGFVLTSPKNPIKSVTKFVA
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    YNETETEVDMAMCGEQVKLRIKGVEEDDISPGFVLTSPKNPIKNVTKFVA
                                             **** *****:*****:**:*****::**.**********:*.**:***

SUP35_Kla_AB039749_1                        QIAIVELKSIMSAGFSCVMHIHTAIEEVTVTRLLHKLEKGSNRKSKKPPA
SUP35_Agos_ATCC_10895_NM_211584_1           QIAIVELKSIMSAGFSCVMHVHTAIEEVSITRLLHKLEKGTNRKSKKPPA
SUP35_Scer_74-D694_GCA_001578265.1_1        QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Spar_A12_Liti_1                       QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Smik_IFO1815T_30_1                    QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Skud_IFO1802T_36_1                    QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGSNRKSKKPPA
SUP35_Sbou_unique28_CM003560_1              QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Scer_beer078_CM005938_1               QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Sarb_H-6_chrXIII_CM001575_1           QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    QIAIVELKSIIAAGFSCVMHVHTAIEEVHIVKLLHKLEKGTNRKSKKPPA
                                            **********::********:******* :.:*********:*********

SUP35_Kla_AB039749_1                        FAKKGMKIIAVIETNEPVCVETYDDYPQLGRFTLRDQGTTIAIGKIVKIL
SUP35_Agos_ATCC_10895_NM_211584_1           FAKKGMKIIAVLETEEPVCVETYQDYPHLGRFTLRDQGITIAIGKIVKIL
SUP35_Scer_74-D694_GCA_001578265.1_1        FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Spar_A12_Liti_1                       FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Smik_IFO1815T_30_1                    FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGATIAIGKIVKIA
SUP35_Skud_IFO1802T_36_1                    FAKKGMKVIAVLETEAPVCVEAYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Sbou_unique28_CM003560_1              FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Scer_beer078_CM005938_1               FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Sarb_H-6_chrXIII_CM001575_1           FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    FAKKGMKVIAVLETEAPVCVETYQDYPQLGRFTLRDQGTTIAIGKIVKIA
                                            *******:***:**: *****:* *****:*:***:********** **********

SUP35_Kla_AB039749_1                        EN
SUP35_Agos_ATCC_10895_NM_211584_1           E-
SUP35_Scer_74-D694_GCA_001578265.1_1        E-
SUP35_Spar_A12_Liti_1                       E-
SUP35_Smik_IFO1815T_30_1                    E-
SUP35_Skud_IFO1802T_36_1                    E-
SUP35_Sbou_unique28_CM003560_1              E-
SUP35_Scer_beer078_CM005938_1               E-
SUP35_Sarb_H-6_chrXIII_CM001575_1           E-
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1    E-
                                            *




OUTPUT RESULTS
        #### File Type= GUIDE_TREE      Format= newick          Name= SUP35_10seqs.g.dnd
        #### File Type= MSA             Format= fasta_aln       Name= SUP35_10seqs_tcoffee.faa

# Command Line: /home/isemenov/anaconda3/lib/t_coffee-11.0.8/bin/t_coffee -infile SUP35_10seqs.g.faa
-outfile SUP35_10seqs_tcoffee.faa -output fasta_aln  [PROGRAM:T-COFFEE]
# T-COFFEE Memory Usage: Current= 30.926 Mb, Max= 34.044 Mb
# Results Produced with T-COFFEE Version_11.00.8cbe486 (2014-08-12 22:05:29 - Revision 8cbe486 - Buil
d 477)
# T-COFFEE is available from http://www.tcoffee.org
# Register on: https://groups.google.com/group/tcoffee/

*****************************************************************************
Result: None
Error message: None
*****************************************************************************

32.24user 0.42system 0:05.24elapsed 622%CPU (0avgtext+0avgdata 119936maxresident)k
12912inputs+9048outputs (28major+188267minor)pagefaults 0swaps
```

In [47]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_tcoffee.faa", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1
  752
SUP35_Agos_ATCC_10895_NM_211584_1
  752
SUP35_Scer_74-D694_GCA_001578265.1_1
  752
SUP35_Spar_A12_Liti_1
  752
SUP35_Smik_IFO1815T_30_1
  752
SUP35_Skud_IFO1802T_36_1
  752
SUP35_Sbou_unique28_CM003560_1
  752
SUP35_Scer_beer078_CM005938_1
  752
SUP35_Sarb_H-6_chrXIII_CM001575_1
  752
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1
  752
```

Prank

In [48]:
```
! time prank -d=SUP35_10seqs.g.faa -o=SUP35_10seqs_prank.faa
```

```
-----------------
 PRANK v.170427:
-----------------

Input for the analysis
 - aligning sequences in 'SUP35_10seqs.g.faa'
 - using inferred alignment guide tree
 - option '+F' is not used; it can be enabled with '+F'
 - external tools available:
    MAFFT for initial alignment
    Exonerate for alignment anchoring

Warning: sequence names changed.

Correcting (arbitrarily) for multifurcating nodes.
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 1.

Alignment score: 1285
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 2.

Alignment score: 1297
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 3.

Alignment score: 1292
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 4.

Alignment score: 1292
Correcting (arbitrarily) for multifurcating nodes.

Generating multiple alignment: iteration 5.

Alignment score: 1292


Writing
 - alignment to 'SUP35_10seqs_prank.faa.best.fas'

Analysis done. Total time 9s

8.86user 0.14system 0:09.03elapsed 99%CPU (0avgtext+0avgdata 23516maxresident)k
0inputs+416outputs (1major+97565minor)pagefaults 0swaps
```

In [49]:
```
record_dict = IO.to_dict(IO.parse("SUP35_10seqs_prank.faa.best.fas", "fasta"))
for key in record_dict.items():
    print(key[0],"\n ", len(key[1].seq))
```

```
SUP35_Kla_AB039749_1_[1_-_2100]
  791
SUP35_Agos_ATCC_10895_NM_211584_1_[1_-_2073]
  791
SUP35_Scer_74-D694_GCA_001578265.1_1_[1_-_2055]
  791
SUP35_Sbou_unique28_CM003560_1_[1_-_2055]
  791
SUP35_Scer_beer078_CM005938_1_[1_-_1998]
  791
SUP35_Spar_A12_Liti_1_[1_-_2043]
  791
SUP35_Smik_IFO1815T_30_1_[1_-_2043]
  791
SUP35_Sarb_H-6_chrXIII_CM001575_1_[1_-_2040]
  791
SUP35_Skud_IFO1802T_36_1_[1_-_2034]
  791
SUP35_Seub_CBS12357_chr_II_IV_DF968535_1_[1_-_2022]
  791
```

Compare algorithms on 10 protein sequences

algorithm	alignment length (nt)	time (seconds)
ClustalW	719	0.38
CLustalO	757	1.22
MUSCLE	743	0.16
MAFFT	759	0.37
Kalign	720	0.04
t-coffee	752	31.96
prank	791	9.91

Conslusion: the best algorithm here is **MAFFT**. It has rather good length of alignment and speed (balance between these two metrics)

Adding sequences to alignments

MAFFT and Muscle allows to add new sequences to existing alignments without their recalculations

In [50]:
```
! cat SUP35_2addseqs.fa
```

```
>SUP35_Scer_beer090_CM005746
ATGTCGGATTCAAACCAAGGCAACAATCAGCAAAACTACCAGCAATACAGCCAGAACGGTAACCAACAAC
AAGGTAACAACAGATACCAAGGTTATCAAGCTTACAATGCTCAAGCCCAACCTGCAGGTGGGTACTACCA
AAATTACCAAGGTTATTCTGGGTACCAACAAGGTGGCTATCAACAGTACAATCCCGACGCCGGTTACCAG
CAACAGTATAATCCTCAAGGAGGCTATCAACAGTACAATCCTCAAGGCGGTTATCAGCAGCAATTCAATC
CACAAGGTGGCCGTGGAAATTACAAAAACTTCAACTACAATAACAATTTGCAAGGATATCAAGCTGGTTT
CCAACCACAGTCTCAAGGTATGTCTTTGAACGACTTTCAAAAGCAACAAAAGCAGGCCGCTCCCAAACCA
AAGAAGACTTTGAAGCTTGTCTCCAGTTCCGGTATCAAGTTGGCCAATGCTACCAAGAAGGTTGACACAA
AACCTGCCGAATCTGATAAGAAAGAGGAAGAGAAGTCTGCTGAAACCAAAGAACCAACTAAAGAGCCAAC
AAAGGTCGAAGAACCAGTTAAAAAGGAGGAGAAACCAGTCCAGACTGAAGAAAAGACGGAGGAAAAAATCG
GAACTTCCAAAGGTAGAAGACCTTAAAATCTCTGAATCAACACATAATACCAACAATGCCAATGTTACCA
GTGCTGATGCCTTGATCAAGGAACAGGAAGAAGAAGTGGATGACGAAGTTGTTAACGATATGTTTGGTGG
TAAAGATCACGTTTCTTTAATTTTCATGGGTCATGTTGATGCCGGTAAATCTACTATGGGTGGTAATCTA
CTATACTTGACTGGCTCTGTGGATAAGAGAACTATTGAGAAATATGAAAGAGAAGCCAAGGATGCAGGCA
GACAAGGTTGGTACTTGTCATGGGTCATGGATACCAACAAAGAAGAAAGAAATGATGGTAAGACTATCGA
AGTTGGTAAGGCCTACTTTGAAACTGAAAAAAGGCGTTATACCATATTGGATGCTCCTGGTCATAAAATG
TACGTTTCCGAGATGATCGGTGGTGCTTCTCAAGCTGATGTTGGTGTTTTGGTCATTTCCGCCAGAAAGG
GTGAGTACGAAACCGGTTTTGAGAGAGGTGGTCAAACTCGTGAACACGCCCTATTGGCCAAGACCCAAGG
TGTTAATAAGATGGTTGTCGTCGTAAATAAGATGGATGACCCAACCGTTAACTGGTCTAAGGAACGTTAC
GACCAATGTGTGAGTAATGTCAGCAATTTCTTGAGAGCAATTGGTTACAACATTAAGACAGACGTTGTAT
TTATGCCAGTATCCGGCTACAGTGGTGCAAATTTGAAAGATCACGTAGATCCAAAAGAATGCCCATGGTA
CACCGGCCCAACTCTGTTAGAATATCTGGATACAATGAACCACGTCGACCGTCACATCAATGCTCCATTC
ATGTTGCCTATTGCCGCTAAGATGAAGGATCTAGGTACCATCGTTGAAGGTAAAATTGAATCCGGTCATA
TCAAAAAGGGTCAATCCACCCTACTGATGCCTAACAAAACCGCTGTGGAAATTCAAAATATTTACAACGA
AACTGAAAATGAAGTTGATATGGCTATGTGTGGTGAGCAAGTTAAACTAAGAATCAAGGGTGTTGAAGAA
GAAGACATTTCACCAGGTTTTGTACTAACATCGCCAAAGAACCCTATCAAGAGTGTTACCAAGTTTGTAG
CTCAAATTGCTATTGTAGAATTAAAATCTATCATAGCAGCCGGTTTTTCATGTGTTATGCATGTTCATAC
AGCAATTGAAGAGGTACATATTGTTAAGTTATTGCACAAATTAGAAAAGGGTACCAACCGTAAGTCAAAG
AAACCACCTGCTTTTGCTAAGAAGGGTATGAAGGTCATCGCTGTTTTAGAAACTGAAGCTCCAGTTTGTG
TGGAAACTTACCAAGATTACCCTCAATTAGGTAGATTCACTTTGAGAGATCAAGGTACCACAATAGCAAT
TGGTAAAATTGTTAAAATTGCCGAGTAA
>SUP35_Scer_EC1118_FN393063.1
ATGTCGGATTCAAACCAAGGCAACAATCAGCAAAACTACCAGCAATACAGCCAGAACGGTAACCAACAAC
AAGGTAACAACAGATACCAAGGTTATCAAGCTTACAATGCTCAAGCCCAACCTGCAGGTGGGTACTACCA
AAATTACCAAGGTTATTCTGGGTACCAACAAGGTGGCTATCAACAGTACAATCCCGACGCCGGTTACCAG
CAACAGTATAATCCTCAAGGAGGCTATCAACAGTACAATCCTCAAGGCGGTTATCAGCAGCAATTCAATC
CACAAGGTGGCCGTGGAAATTACAAAAACTTCAACTACAATAACAATTTGCAAGGATATCAAGCTGGTTT
CCAACCACAGTCTCAAGGTATGTCTTTGAACGACTTTCAAAAGCAACAAAAGCAGGCCGCTCCCAAACCA
AAGAAGACTTTGAAGCTTGTCTCCAGTTCCGGTATCAAGTTGGCCAATGCTACCAAGAAGGTTGACACAA
AACCTGCCGAATCTGAAAAGAAAGAGGAAGAGAAGTCTGCTGAAACCAAAGAACCAACTAAAGAGCCAAC
AAAGGTCGAAGAACCAGTTAAAAAGGAGGAGAAACCAGTCCAGACTGAAGAAAAGAAGGAGGAAAAAATCG
GAACTTCCAAAGGTAGAAGACCTCAAAATCTCTGAATCAACACATAATACCAACAATGCCAATGTTACCA
GTGCTGATGCCTTGATCAAGGAACAGGAAGAAGAAGTGGATGACGAAGTTGTTAACGATATGTTTGGTGG
TAAAGATCACGTTTCTTTAATTTTCATGGGTCATGTTGATGCCGGTAAATCTACTATGGGTGGTAATCTA
CTATACTTGACTGGCTCTGTGGATAAGAGAACTATTGAGAAATATGAAAGAGAAGCCAAGGATGCAGGCA
GACAAGGTTGGTACTTGTCATGGGTCATGGATACCAACAAAGAAGAAAGAAATGATGGTAAGACTATCGA
AGTTGGTAAGGCCTACTTTGAAACTGAAAAAAGGCGTTATACCATATTGGATGCTCCTGGTCATAAAATG
TACGTTTCCGAGATGATCGGTGGTGCTTCTCAAGCTGATGTTGGTGTTTTGGTCATTTCCGCCAGAAAGG
GTGAGTACGAAACCGGTTTTGAGAGAGGTGGTCAAACTCGTGAACACGCCCTATTGGCCAAGACCCAAGG
TGTTAATAAGATGGCTGTCGTCGTAAATAAGATGGATGACCCAACCGTTAACTGGTCTAAGGAACGTTAC
GACCAATGTGTGAGTAATGTCAGCAATTTCTTGAGAGCAATTGGTTACAACATTAAGACAGACGTTGTAT
TTATGCCAGTATCCGGCTACAGTGGTGCAAATTTGAAAGATCACGTAGATCCAAAAGAATGCCCATGGTA
CACCGGCCCAACTCTGTTAGAATATCTGGATACAATGAACCACGTCGACCGTCACATCAATGCTCCATTC
ATGTTGCCTATTGCCGCTAAGATGAAGGATCTAGGTACCATCGTTGAAGGTAAAATTGAATCCGGTCATA
TCAAAAAGGGTCAATCCACCCTACTGATGCCTAACAAAACCGCTGTGGAAATTCAAAATATTTACAACGA
AACTGAAAATGAAGTTGATATGGCTATGTGTGGTGAGCAAGTTAAACTAAGAATCAAGGGTGTTGAAGAA
GAAGACATTTCACCAGGTTTTGTACTAACATCGCCAAAGAACCCTATCAAGAGTGTTACCAAGTTTGTAG
CTCAAATTGCTATTGTAGAATTAAAATCTATCATAGCAGCCGGTTTTTCATGTGTTATGCATGTTCATAC
AGCAATTGAAGAGGTACATATTGTTAAGTTATTGCACAAATTAGAAAAGGGTACCAACCGTAAGTCAAAG
AAACCACCTGCTTTTGCTAAGAAGGGTATGAAGGTCATCGCTGTTTTAGAAACTGAAGCTCCAGTTTGTG
TGGAAACTTACCAAGATTACCCTCAATTAGGTAGATTCACTTTGAGAGATCAAGGTACCACAATAGCAAT
TGGTAAAATTGTTAAAATTGCCGAGTAA
```

Firstly, we align these two sequences to each other

In [51]:
```
! muscle -in SUP35_2addseqs.fa -out SUP35_2addseqs_muscle.fa
```

```
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

SUP35_2addseqs 2 seqs, lengths min 2058, max 2058, avg 2058
00:00:00    15 MB(-2%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    15 MB(-2%)  Iter   1  100.00%  K-mer dist pass 2
00:00:00    27 MB(-3%)  Iter   1  100.00%  Align node
00:00:00    27 MB(-3%)  Iter   1  100.00%  Root alignment
```

And then add them to existing alignments of our 250 sequences

In [52]:
```
! muscle -profile -in1 SUP35_250seqs_muscle.fa -in2 SUP35_2addseqs.fa -out SUP35_252seqs_muscle.fa
```

```
MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

00:00:00    15 MB(-2%)  Reading SUP35_250seqs_muscle.fa
00:00:00    17 MB(-2%)  250 seqs 2313 cols
00:00:00    17 MB(-2%)  Reading SUP35_2addseqs.fa
00:00:00    17 MB(-2%)  2 seqs 2058 cols
00:00:01    18 MB(-2%)  Aligning profiles
00:00:01    29 MB(-4%)  Building output
00:00:01    30 MB(-4%)  Writing output
```

Similiar procedure can be done with MAFFT

In [53]:
```
! mafft --auto SUP35_2addseqs.fa > SUP35_2addseqs_mafft.fa
```

```
outputhat23=16
treein = 0
compacttree = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
All-to-all alignment.
tbfast-pair (nuc) Version 7.453
alg=L, model=DNA200 (2), 2.00 (6.00), -0.10 (-0.30), noshift, amax=0.0
0 thread(s)

outputhat23=16
Loading 'hat3.seed' ...
done.
Writing hat3 for iterative refinement
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00
tbutree = 1, compacttree = 0
Constructing a UPGMA tree ...
    0 / 2
done.

Progressive alignment ...
STEP     1 /1
done.
tbfast (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
1 thread(s)

minimumweight = 0.000010
autosubalignment = 0.000000
nthread = 0
randomseed = 0
blosum 62 / kimura 200
poffset = 0
niter = 16
sueff_global = 0.100000
nadd = 16
Loading 'hat3' ... done.
generating a scoring matrix for nucleotide (dist=200) ... done

    0 / 2
Segment   1/  1    1-2059
done 001-001-1  identical.
dvtditr (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 L-INS-i (Probably most accurate, very slow)
 Iterative refinement method (<16) with LOCAL pairwise alignment information

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.
```

In [54]:
```
! mafft --add SUP35_2addseqs_mafft.fa SUP35_250seqs_mafft.fa > SUP35_252seqs_mafft.fa
```

```
nadd = 2
nthread = 0
nthreadpair = 0
nthreadtb = 0
ppenalty_ex = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00



Making a distance matrix ..

There are 12 ambiguous characters.
  201 / 252
done.

Constructing a UPGMA tree (efffree=0) ...
  250 / 252
done.

Progressive alignment 1/2...
STEP   208 / 251
done.

Making a distance matrix from msa..
  200 / 252
done.

Constructing a UPGMA tree (efffree=1) ...
  250 / 252
done.

Progressive alignment 2/2...
STEP   209 / 251
done.

disttbfast (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-2 (Fast but rough)
 Progressive method (guide trees were built 2 times.)

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.
```