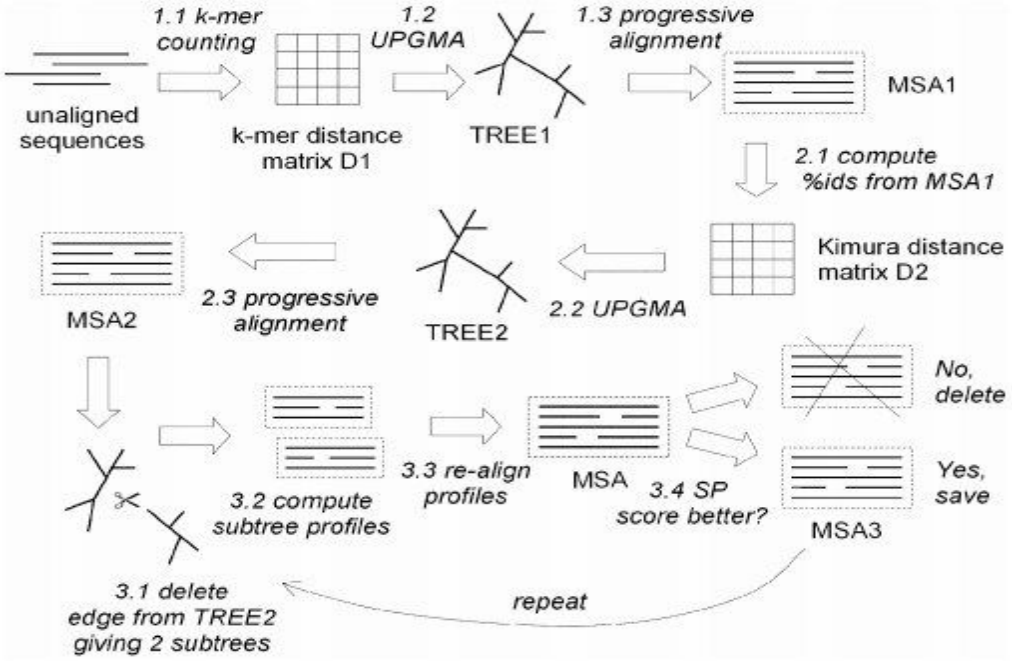
A comprehensive performance analysis of various MSA tools

Arnav Aima - 99864675
Sakshi Dubey - 48131141
Suhani Mehta - 47986909

TOOLS ANALYZED

- **MUSCLE**
- **Kalign**
- **T-Coffee**

MUSCLE


1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
3.1 delete edge from TREE2 giving 2 subtrees
repeat

MUSCLE

Three stages:

Draft Progressive

Improved Progressive

Refinement: multiple alignments at each stage

Kalign

- Progressive alignment algorithm
- Does multi-pattern matching and global DP
- Uses approximate string matching using Wu-Manber

T-Coffee

- MSA software using progressive approach.
- Generates a library of pairwise alignments.
- Allows heterogeneous comparison of alignment.

Dataset – BAliBASE

Reference Set	No of Seq Sets	Reference Set Description
RV11	38	Equidistant sequences (<20% identity)
RV12	44	Equidistant sequences (20%-40% identity)
RV20	41	Highly divergent orphan sequences
RV30	30	Subgroups with <25% residue identity
RV40	49	N/C-terminal extensions
RV50	16	Internal insertions

Evaluation Metrics

- Q Score
- TC Score
- Cline Score
- Running Time

Helper Tools

- Qscore
- Seqret
- Rose

Screenshots

QScore Comparison

```
Arnavs-MacBook-Pro:qscore_src arnavaima$ javac Testing.java
Arnavs-MacBook-Pro:qscore_src arnavaima$ java Testing
Test=BB50001OUT.tfa;Ref=BB50001IN.tfa;Q=0.721;TC=0.317;Cline=0.699
Test=BB50002OUT.tfa;Ref=BB50002IN.tfa;Q=0.187;TC=0;Cline=0.0808
Test=BB50003OUT.tfa;Ref=BB50003IN.tfa;Q=0.574;TC=0.2;Cline=0.544
Test=BB50004OUT.tfa;Ref=BB50004IN.tfa;Q=0.938;TC=0.828;Cline=0.947
Test=BB50005OUT.tfa;Ref=BB50005IN.tfa;Q=0.918;TC=0.707;Cline=0.926
Test=BB50006OUT.tfa;Ref=BB50006IN.tfa;Q=0.55;TC=0.0715;Cline=0.571
Test=BB50007OUT.tfa;Ref=BB50007IN.tfa;Q=0.625;TC=0.106;Cline=0.689
Test=BB50008OUT.tfa;Ref=BB50008IN.tfa;Q=0.754;TC=0.382;Cline=0.773
Test=BB50009OUT.tfa;Ref=BB50009IN.tfa;Q=0.656;TC=0.063;Cline=0.606
Test=BB50010OUT.tfa;Ref=BB50010IN.tfa;Q=0.516;TC=0.0497;Cline=0.409
Test=BB50011OUT.tfa;Ref=BB50011IN.tfa;Q=0.597;TC=0.112;Cline=0.596
Test=BB50012OUT.tfa;Ref=BB50012IN.tfa;Q=0.516;TC=0.0367;Cline=0.572
Test=BB50013OUT.tfa;Ref=BB50013IN.tfa;Q=0.885;TC=0.63;Cline=0.887
Test=BB50014OUT.tfa;Ref=BB50014IN.tfa;Q=0.793;TC=0.349;Cline=0.823
Test=BB50015OUT.tfa;Ref=BB50015IN.tfa;Q=0.537;TC=0.039;Cline=0.597
Test=BB50016OUT.tfa;Ref=BB50016IN.tfa;Q=0.521;TC=0.0981;Cline=0.547
Average QScore: 0.643; Average TC Score: 0.2493125; Average Cline Score: 0.641675
Arnavs-MacBook-Pro:qscore_src arnavaima$
```

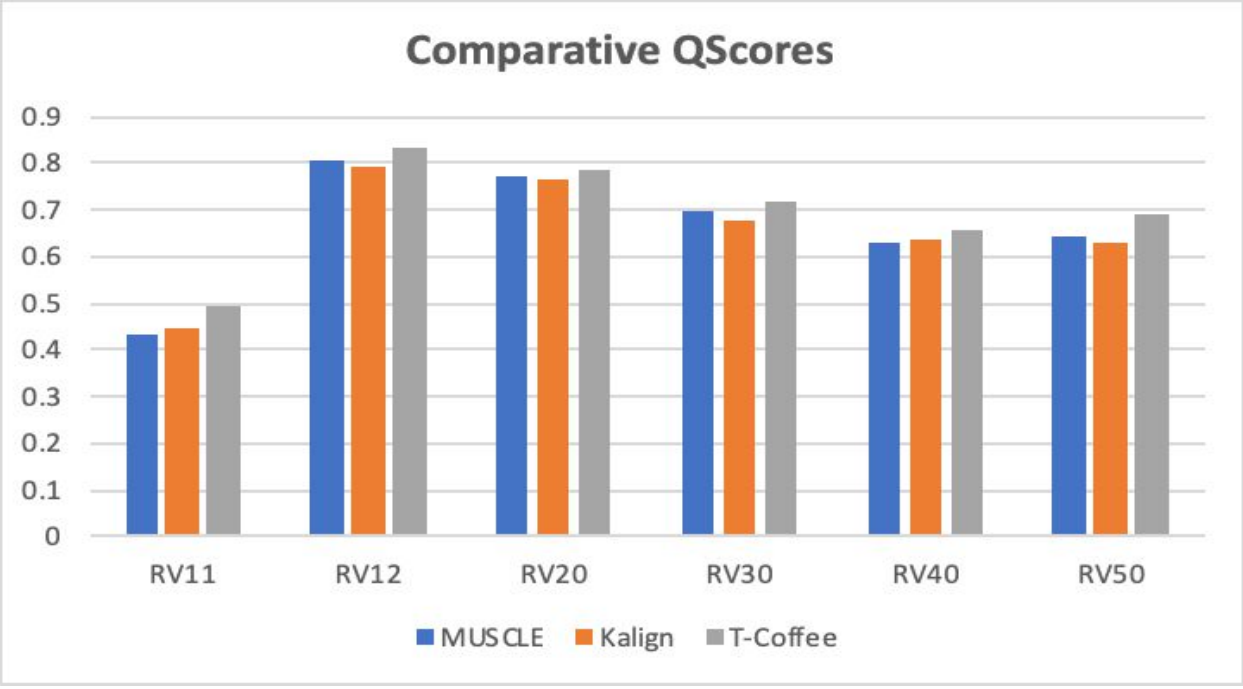
Kalign

```
sumehta@DESKTOP-02INI4V:/mnt/c/Users/suhan/Documents/BioInformatics/project/current$ kalign ../RV40/BB40001.tfa ../KalignBB40001.tfa
EXTRA :2

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 no sequences.

reading from ../RV40/BB40001.tfa: found 28 sequences
reading from ../KalignBB40001.tfa: found no sequences.
-> output file, in fasta format

Aligning 28 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 file:'../KalignBB40001.tfa'.

Distance Calculation:
     100 percent done
Alignment:
     100 percent done
```

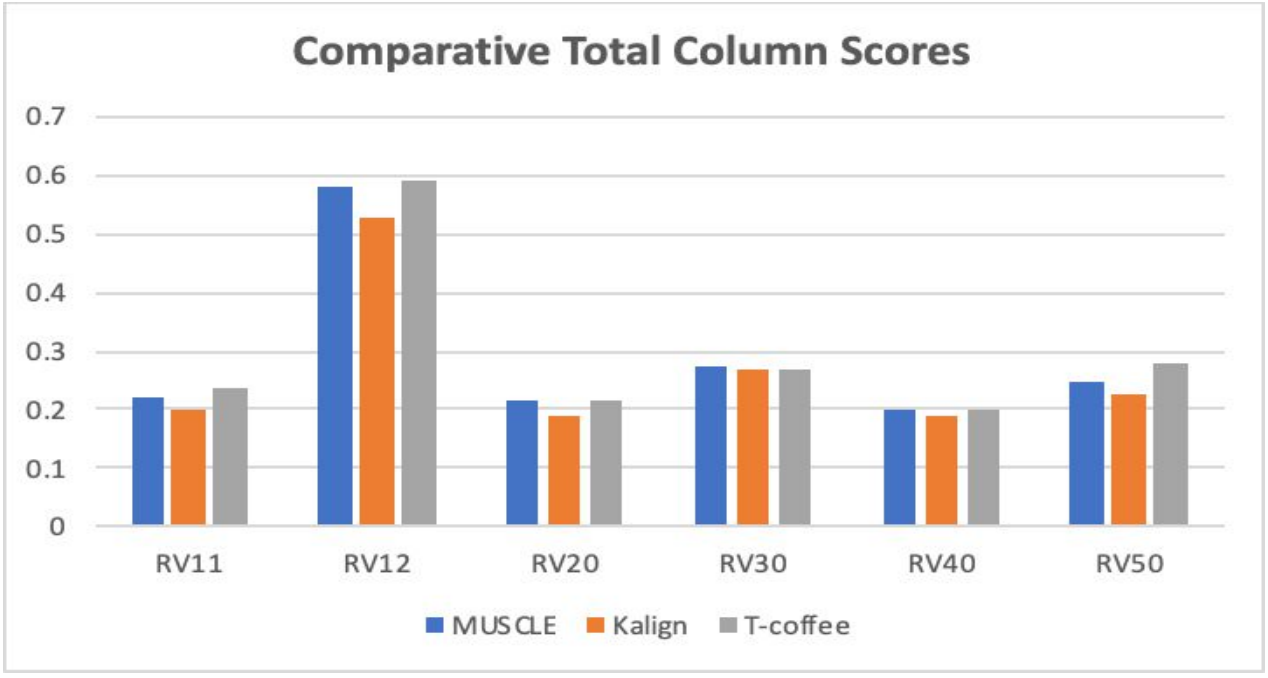
Dataset – BAliBASE

Reference Set	No of Seq Sets	Reference Set Description
RV11	38	Equidistant sequences (<20% identity)
RV12	44	Equidistant sequences (20%-40% identity)
RV20	41	Highly divergent orphan sequences
RV30	30	Subgroups with <25% residue identity
RV40	49	N/C-terminal extensions
RV50	16	Internal insertions

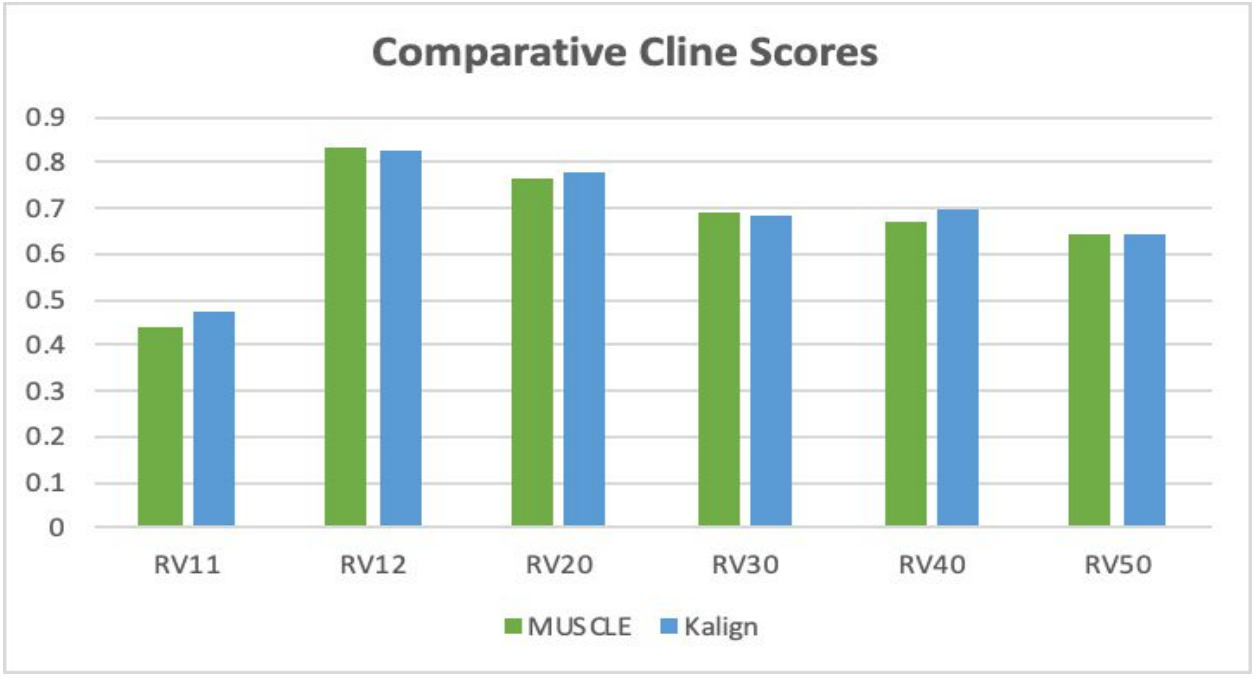
Q Score


Comparative QScores
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RV11
RV12
RV20
RV30
RV40
RV50
MUSCLE
Kalign
T-Coffee

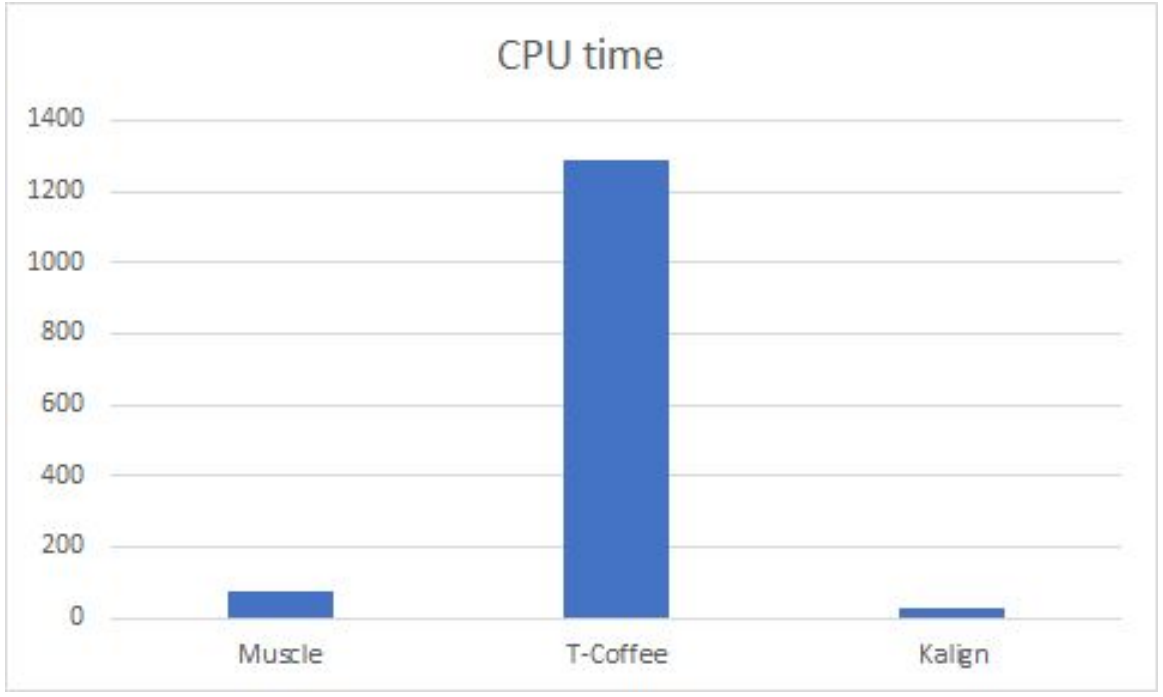
TC Score


Comparative Total Column Scores
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RV11
RV12
RV20
RV30
RV40
RV50
MUSCLE
Kalign
T-coffee

Cline Score


Comparative Cline Scores
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
RV11
RV12
RV20
RV30
RV40
RV50
MUSCLE
Kalign

Run Time


CPU time
1400
1200
1000
800
600
400
200
0
Muscle
T-Coffee
Kalign

Conclusion

Tool	Accuracy	Speed
MUSCLE	MODERATE	FAST
Kalign	MODERATE	VERY FAST
T-Coffee	HIGH	VERY SLOW

Kalign has the perfect combination of accuracy and speed which is mostly due to its extremely efficient string matching algorithm (Wu-Manber) and is the most suitable tool for MSA as per our observations.

Thank You

Questions?