Use of parallelism on MSA tools

Miquel Orobitg Cortada

September 9, 2011

Index

1. Multiple Sequence Alignment
2. MSA Tools
   - T-Coffee
   - Parallel-TCoffee
3. Proposed Solutions
   - Balanced Guide Tree
   - Multiple Trees
4. Future Work

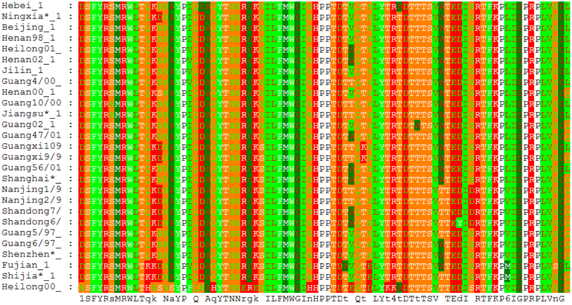
Index

1. Multiple Sequence Alignment
2. MSA Tools
   - T-Coffee
   - Parallel-TCoffee
3. Proposed Solutions
   - Balanced Guide Tree
   - Multiple Trees
4. Future Work

Sequence Alignment

Definition

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.



```
Hebei_1     : DSFYRSMRWLTQKNNAYPIQGAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRADTTTSVATEDINRTFKPLIGPRPLVNGL
Ningxia*_1  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Beijing_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan98_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Heilong01_  : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan02_1   : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Jilin_1     : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guang4/00_  : DSFYRSMRWLTQKNNAYPVQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Henan00_1   : DSFYRSMRWLTQKSNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guang10/00  : DSFYRNMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Jiangsu*_1  : DSFYRNMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTTQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guang02_1   : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTVQTNLYTRTDTATSVATEDINRTFKPLIGPRPLVNGQ
Guang47/01  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKNILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guangxi109  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQKNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGL
Guangxi9/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQKNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Guang56/01  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVATEEINRTFKPLIGPRPLVNGL
Shanghai*_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVATEDINRTFKPLIGPRPLVNGQ
Nanjing1/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Nanjing2/9  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTTQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Shandong7/  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVTTEDIDRTFKPVIGPRPLVNGQ
Shandong6/  : DSFYRSMRWLTQKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTAQTNLYTRTDTTTSVTTEYIDRTFKPVIGPRPLVNGQ
Guang5/97_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVTTEDINRTFKPVIGPRPLVNGQ
Guang6/97_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPVIGPRPLVNGQ
Shenzhen*_  : DSFYRSMRWLTQKNNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTNLYTRTDTTTSVATEDINRTFKPVIGPRPLVNGQ
Fujian_1    : DSFYRSMRWLTKKDNAYPVQDAQYTNNRGKSILFMWGINHPPTDAVQTELYTRTDTTTSVATEDINRTFKPMIGPRPLVSGL
Shijia*_1   : DSFYRSMRWLTKKDNAYPIQDAQYTNNRGKSILFMWGINHPPTDTVQTELYTRTDTTTSVTTEDINRTFKPMIGPRPLVNGQ
Heilong00_  : NSFYRSMRWLTHNSNSYPFQNAHYTNNERENILFMWGIHHPPTDTEQTDLYKKADTTTSVTTEDISRTFKPVIGPRPLVNGQ
              1SFYRsMRWLTqk NaYP Q AqYTNNrgk ILFMWGInHPPTDt Qt LYt4tDTtTSV TEdI RTFKP6IGPRPLVnG
```

Sequence Alignment

Types

Types

- Pairwise Alignments.
- Multiple Sequence Alignments (MSA)

Multiple Sequence Alignment

Global Optimization Methods

Dynamic programming

- Is a technique to identify the globally optimal alignment solution.
- Exists different algorithms (global, local, glocal).

Problems

- Computationally difficult to produce the alignment (NP-complete problem).

Sequence Alignment

Heuristics (1/2)

Progressive Alignment (PA)

- The alignment is produced by a successive construction of pair-wise alignments.
- Advantages:
  - Good compromise between time spend and accuracy.
- Disadvantages:
  - Heavy dependence on the initial alignment.
  - It is not guaranteed to converge to a global optimum.
- Common methods: T-Coffee and ClustalW

Sequence Alignment

Heuristics (2/2)

Iterative methods

- Tries to reduce the errors made in progressive methods.
- Works similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA.
- Common methods: Dialign and Muscle

Index

1. Multiple Sequence Alignment
2. MSA Tools
   - T-Coffee
   - Parallel-TCoffee
3. Proposed Solutions
   - Balanced Guide Tree
   - Multiple Trees
4. Future Work

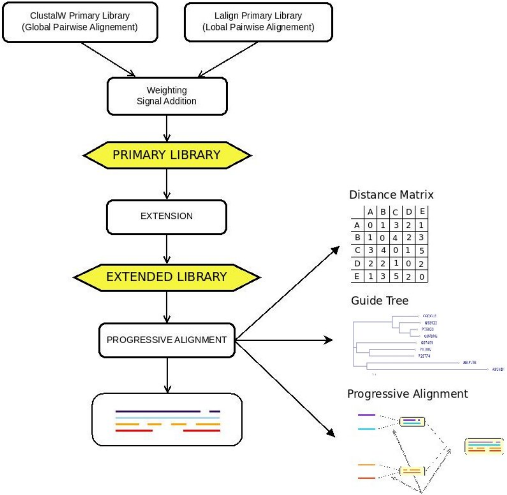
T-Coffee

T-Coffee

- Is a MSA method that combines the consistency based scoring function COFFEE with the progressive alignment algorithm.
- Advantages:
  - Improvement in the accuracy compared with progressive methods.
  - Reduce the dependency on the initial alignment.

T-Coffee

Structure



	A	B	C	D	E
A	0	1	3	2	1
B	1	0	4	2	3
C	3	4	0	1	5
D	2	2	1	0	2
E	1	3	5	2	0

T-Coffee

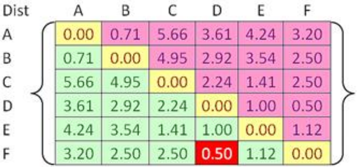
1. Library

Library

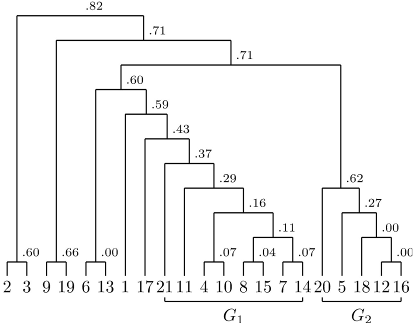
- List of pairs of alignments evaluated by a weight that is given by a percentage of identity.
- Generated using different resources.
- Can be extended by transitive properties.
- Used in the progressive alignment.

Structure

2. Distance Matrix (DM) & Guide Tree (GT)



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00



Structure

3. Progressive Alignment

Progressive Alignment (PA)

- Align the two input sequences using the information of the library.
- The order is determined by the alignment guide tree.

Disadvantages

Library

- $Size = N^2 * L$
- Primary library complexity: $O(N^2L^2)$
- Extended library complexity: $O(N^3L^2)$

Progressive aligment

- Requires n-1 partial multiple alignments using the library. Each alignment can be computation intensive.
- Complexity: $O(NL^2)$

Index

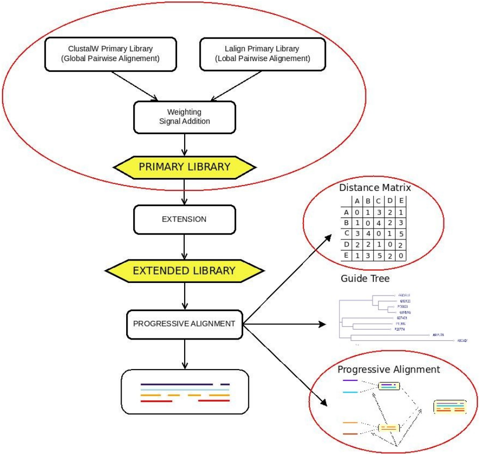
1. Multiple Sequence Alignment
2. MSA Tools
   - T-Coffee
   - Parallel-TCoffee
3. Proposed Solutions
   - Balanced Guide Tree
   - Multiple Trees
4. Future Work

Parallel-TCoffee

Parallelization analysis

3 parallelization steps

1. Library generation
2. Distance Matrix computation
3. Progressive alignment

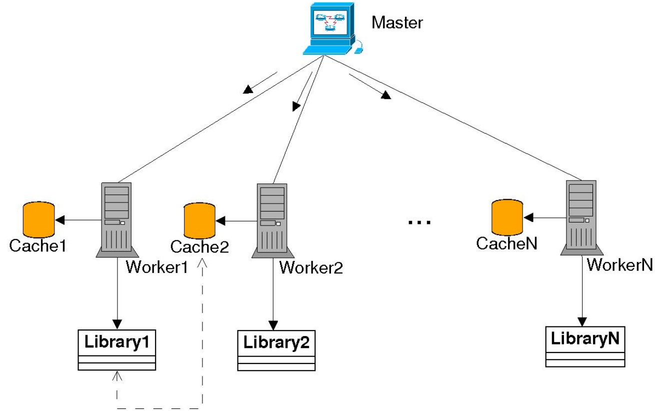


Distance Matrix

	A	B	C	D	E
A	0	1	3	2	1
B	1	0	4	2	3
C	3	4	0	1	5
D	2	2	1	0	2
E	1	3	5	2	0

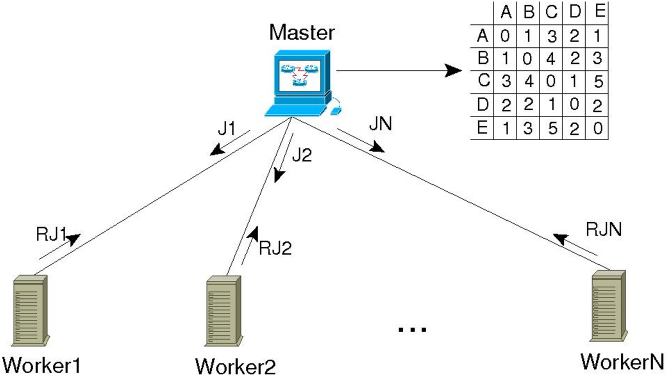
Parallel-TCoffee

1. Library Generation



Parallel-TCoffee

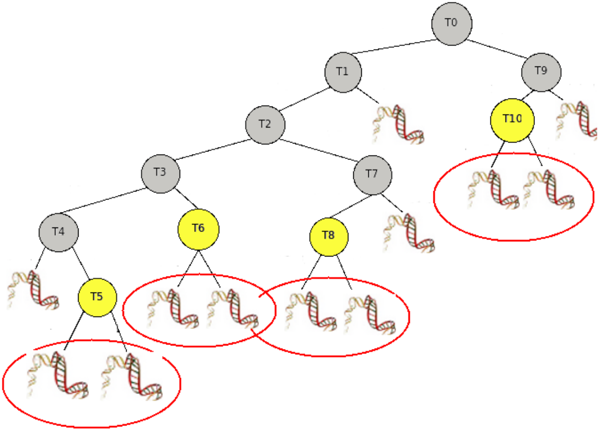
2. Distance Matrix



	A	B	C	D	E
A	0	1	3	2	1
B	1	0	4	2	3
C	3	4	0	1	5
D	2	2	1	0	2
E	1	3	5	2	0

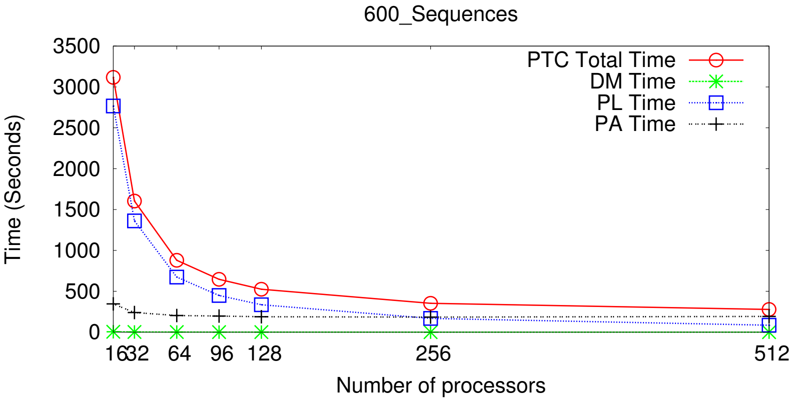
Parallel-TCoffee

3. Progressive Alignment



Parallel-TCoffee Performance

PF00231 Execution times (554 sequences)

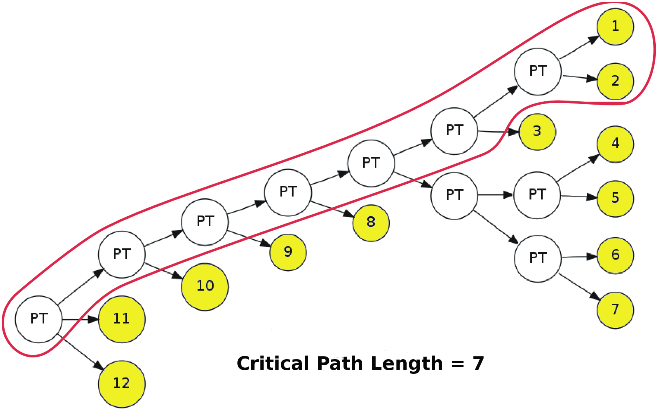


Index

1. Multiple Sequence Alignment
2. MSA Tools
   - T-Coffee
   - Parallel-TCoffee
3. Proposed Solutions
   - Balanced Guide Tree
   - Multiple Trees
4. Future Work

NJ Guide Tree analysis

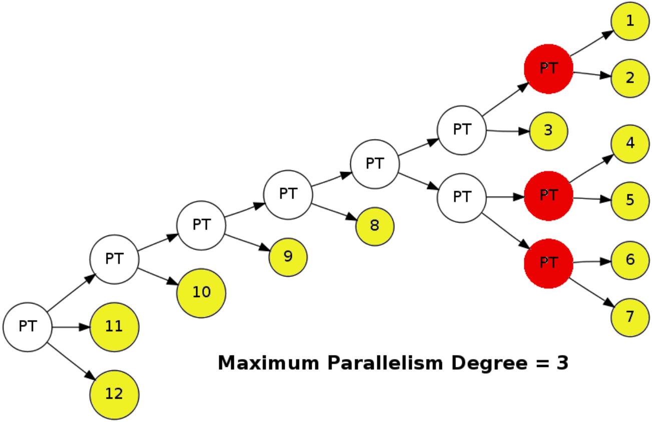
Critical Path (CP)



Critical Path Length = 7

NJ Guide Tree analysis

Maximum Parallelism Degree (MPD)



Maximum Parallelism Degree = 3

NJ Guide Tree analysis

Examples

Sequence Set	Nseqs	CP/OCP	MPD
PF00859	105	37/7	19
PF00074	442	24/9	137
PF00349	515	21/10	144
PF01057	563	84/10	87
PF00007	731	54/10	186

NJ Guide Tree analysis

Guide Tree problems

- Trees generated with T-Coffee are unbalanced.
- Dependence between iterations.
- Low degree of parallelism.
- Limited scalability.

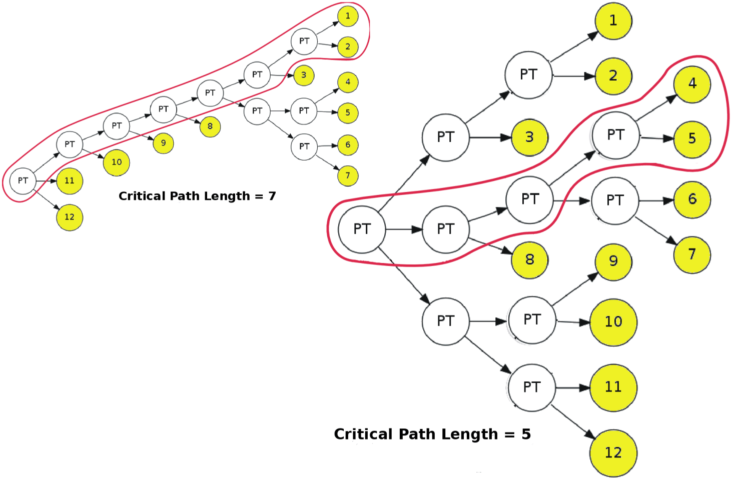
Balanced Guide Tree heuristic

Balanced Guide Tree (BGT)

- BGT: Heuristic to balance the nj guide tree maintaining the alignment accuracy.
- Goals:
  - Reduce the number of precedence relations.
  - Decrease the critical path.
  - Increase the parallelism degree.

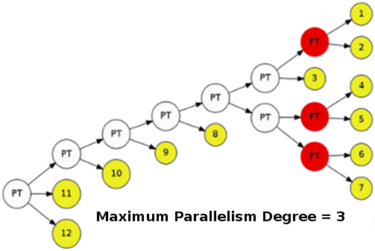
Balanced Guide Tree heuristic

BGT Tree Features - Critical Path (CP)

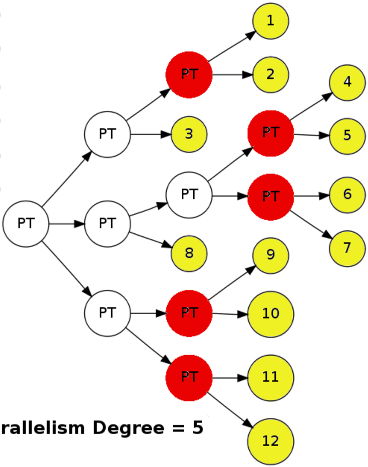


Balanced Guide Tree heuristic

BGT Tree Features - Parallelism Degree (PD)



Maximum Parallelism Degree = 3



Maximum Parallelism Degree = 5

BGT Results

Tree features comparison

	Standard Tree		BGT Tree	
Sequence Set	CP/OCP	MPD	CP/OCP	MPD
PF00859 (105)	37/7	19	8/7	51
PF00074 (442)	24/9	137	14/9	216
PF00349 (515)	21/10	144	15/10	249
PF01057 (563)	84/10	87	17/10	274
PF00007 (731)	54/10	186	24/10	355

BGT Results

Alignment Accuracy

	T-Coffee		BGT	
Balibase	**SP**	**TC**	**SP**	**TC**
Ref 1	0.764	0.579	0.763	0.577
Ref 2	0.877	0.362	0.877	0.363
Ref 3	0.785	0.393	0.783	0.390
Ref 4	0.804	0.419	0,805	0.426
Ref 5	0.788	0.424	0.786	0.426
Ref 6	0,807	0,393	0,807	0,402
Ref 7	0,804	0.360	0.809	0.353
Ref 8	0.700	0.180	0.700	0.180
Ref 9	0.742	0.481	0.742	0.482
Total	**0.783**	**0.457**	**0.783**	**0.458**

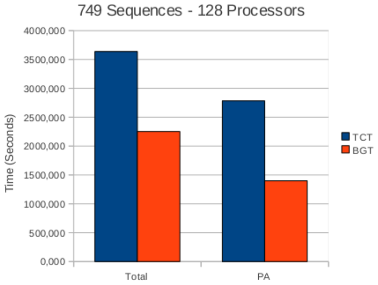
BGT Results

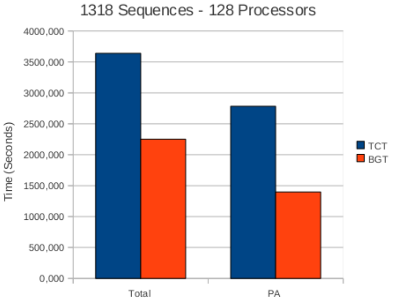
Alignment Accuracy

	T-Coffee	BGT
Prefab	**Q**	**Q**
0 - 15	0.421	0.422
15 - 25	0.724	0.725
25 - 35	0.877	0.875
35 -100	0.955	0.954
Total	**0.711**	**0.711**

BGT Results

Parallel-TCoffee Performance

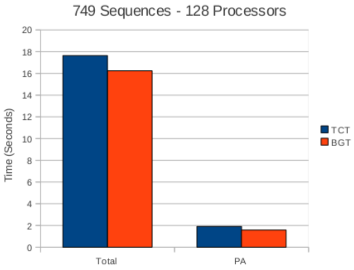


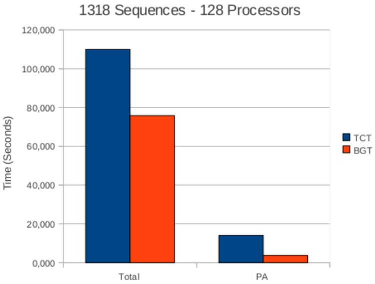


Total: Total execution time
PA: Progressive alignment execution time

BGT Results

ClustalW-MPI Performance

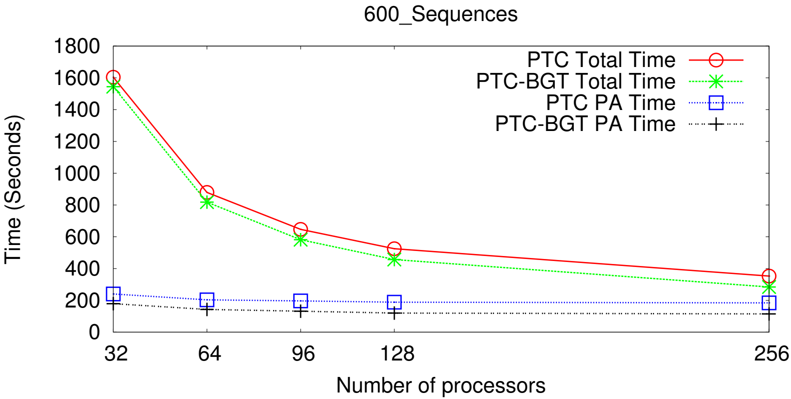




Total: Total execution time
PA: Progressive alignment execution time

BGT Results

Parallel-TCoffee Performance - 600 Sequences



Index

1 Multiple Sequence Alignment

2 MSA Tools
- T-Coffee
- Parallel-TCoffee

3 Proposed Solutions
- Balanced Guide Tree
- Multiple Trees

4 Future Work

Proposal

Proposal

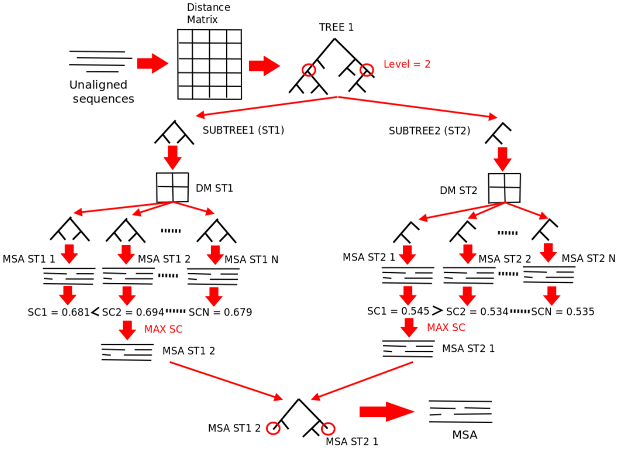
- Create multiple different subtrees of a guide tree.
- Calculate the alignment of each subtree and its score.
- Use the alignment which gets the best score.

Objective

- Improve the alignment accuracy.

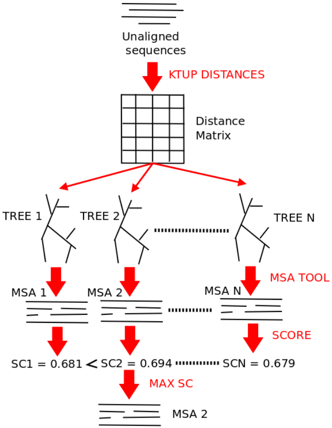
Proposal

Algorithm proposal



First Implementation

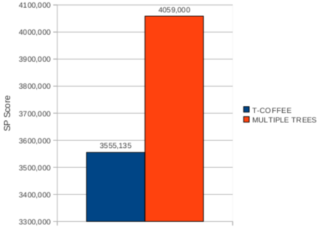
Algorithm

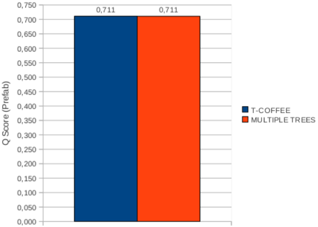


Experimentation results

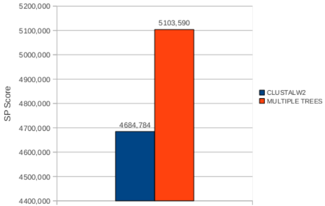
Prefab - SP Score

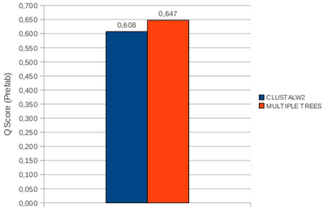
- **T-Coffee:**





- **ClustalW:**

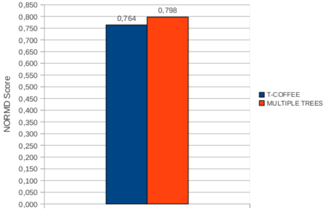


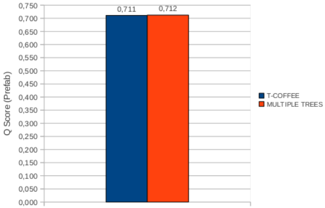


Experimentation results

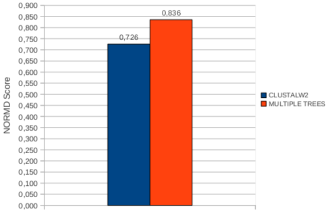
Prefab - NORMD Score

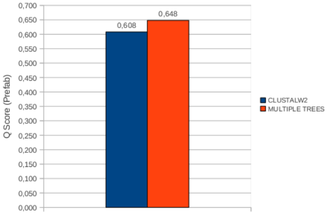
- **T-Coffee:**





- **ClustalW:**

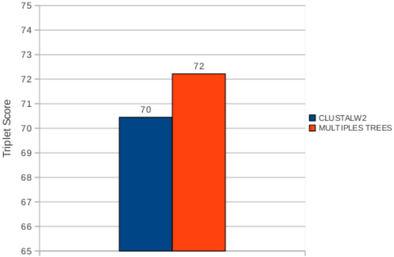


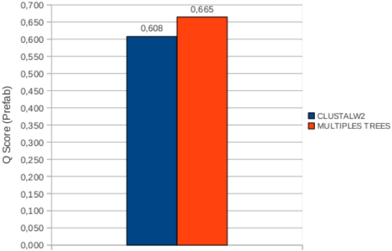


Experimentation results

Prefab - Triplet Score

- **ClustalW:**





Disadvantages

Disadvantages

- To find an evaluation score that defines the best tree.
- The MSA with the best score is not always the best MSA using the benchmark scores.

Index

1. Multiple Sequence Alignment
2. MSA Tools
    - T-Coffee
    - Parallel-TCoffee
3. Proposed Solutions
    - Balanced Guide Tree
    - Multiple Trees
4. Future Work

Future Work

Future Work

- Finish the implementation of the Multiple Trees algorithms.
- Test the performance of the Parallel Multiple Trees solution.
- Publish the Multiple Trees solution.
- Study new parallel algorithms for MSA.

Questions?

Grup de
Computació
Distribuïda

Universitat de Lleida

Universitat de Lleida