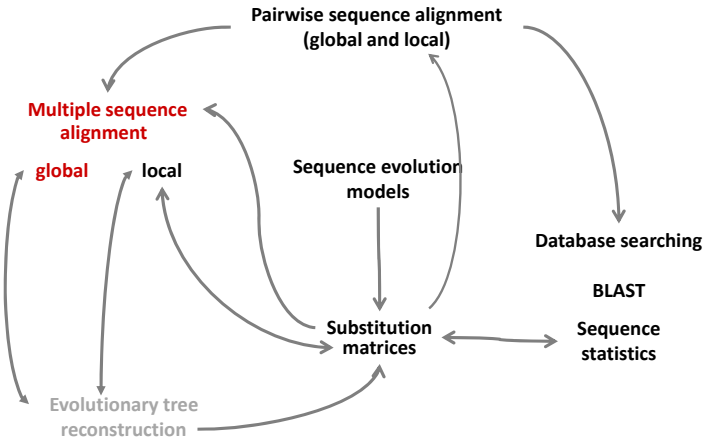


Global Multiple Sequence Alignment

```
HUMAN   MKWVTFISLL FLFSSAYSRG V..FRRDA.H KSEVAHRFKD LGEENFKALV
RABIT   MKWVTFISLL FLFSSAYSRG V..FRREA.H KSEIAHRFND VGEEHFIGLV
PIG     ~~WVTFISLL FLFSSAYSRG V..FRRDT.Y KSEIAHRFKD LGEQYFKGLV
CHICK   MKWVTLISFI FLFSSATSRN LQRFARDAEH KSEIAHRYND LKEETFKAVA
```

Align *k* sequences, so that residues in each column share a property of interest:

- a common ancestor
- a structural or functional role

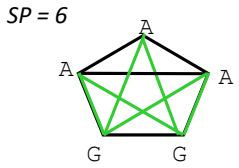
Examples….

Sum-of-pairs scoring

Advantages: Fast, easy to use

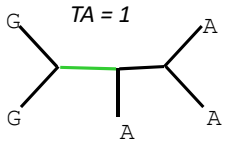
Disadvantages:

- Assumes positional independence
- Over estimates the number of mutations



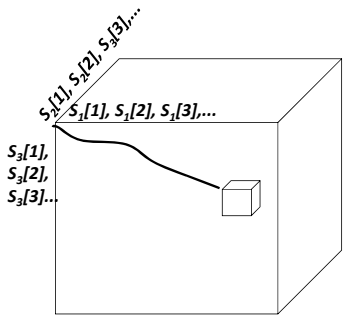
Sum of Pairs

A
A
A
G
G



Tree alignment

Dynamic Programming for Multiple Alignment



Number of cells in matrix: $O(n^k)$

Each cell has $O(2^k)$ neighboring cells

Calculating the sum-of-pairs score for each neighbor is $O(k^2)$

Total computational complexity:

$O(n^k 2^k k^2)$

Limits:

~ k = 8 - 10 sequences

~ n = 500 residues

MSA is NP-complete for Sum-of-Pairs scoring

Observations

1. A multiple alignment induces pairwise alignments
2. A column in the induced pairwise alignment may contain all gaps, even though no column in the MSA contains all gaps.

(1) AG_CT
(2) AG_CT
(3) ACT_T

3. The pairwise alignments induced by the *optimal multiple alignment* are *not* the same as the *optimal pairwise alignments.*

Optimal Pairwise Alignments

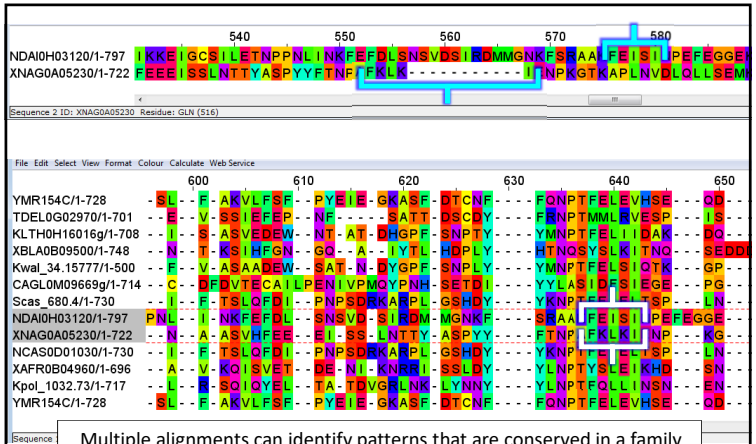
(1) ACT
(2) AGT

1 substitution

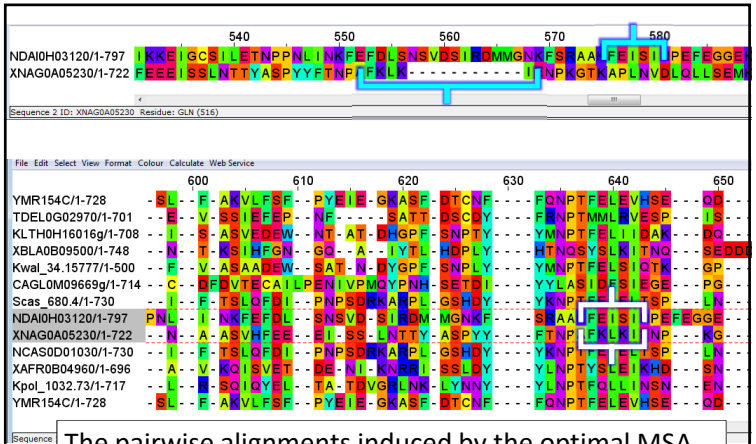
(1) AC_T
(2) A_GT
(3) ACGT

2 indels

Although this costs more, it may be a biologically more realistic alignment



Multiple alignments can identify patterns that are conserved in a family but not apparent in a pairwise alignment of two family members.



The pairwise alignments induced by the optimal MSA

- are not the same as the optimal pairwise alignments,
- are more biologically realistic.

Multiple sequence alignment methods are evaluated against trusted benchmarks

Each benchmark consists of

- A set of sequences from the same protein family
- A reference alignment generated using protein stucture comparison, e.g. BaliBASE, Prefab

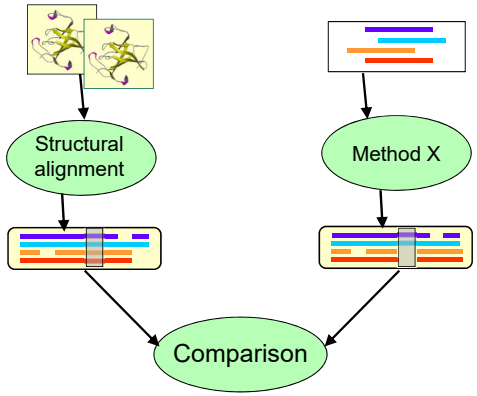
Accuracy is assessed by comparing

- the alignment obtained with Method X with
- the reference alignment

Measures of accuracy:

- the number of correctly aligned columns
- the number of correctly alighed amino acid pairs

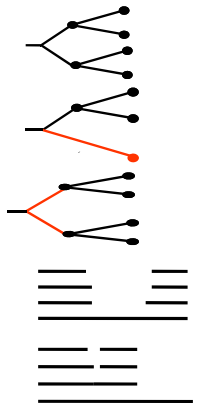
Multiple sequence alignment methods are evaluated against trusted benchmarks



Various benchmarks are designed to mimic properties of different types of data sets encountered in practice, especially those that are challenging to align:

- Highly divergent sequences, e.g., <50% or <30% identity
- A family of related sequences plus several outliers, or “orphan” sequences
- Related sequences that differ due to large N or C terminal extensions or large internal insertions or deletionsAria

Challenging cases

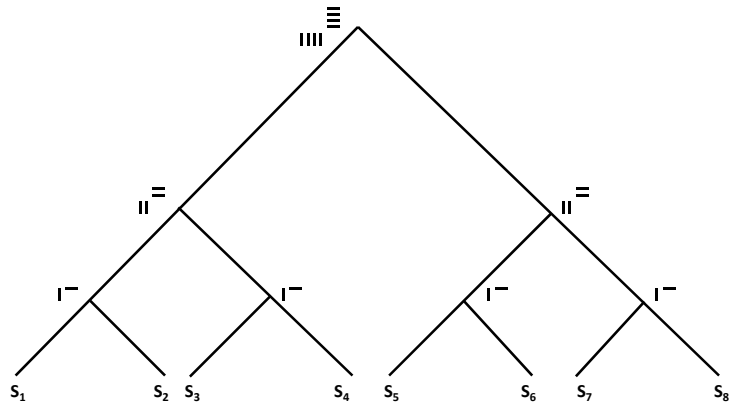


Source: BaliBase, Thompson et al, NAR, 1999,

The progressive alignment heuristic for global multiple sequence alignment

Basic progressive alignment strategy

- Compute *D,* a matrix of distances between all pairs
- Use *D* to construct a "guide tree" *T*
- Construct MSA by
  - pairwise alignment of partial alignments ("profiles")
  - guided by T
  - Using the "once a gap, always a gap" rule.
- Improve alignment by post-processing steps



Optimal Pairwise Alignments

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

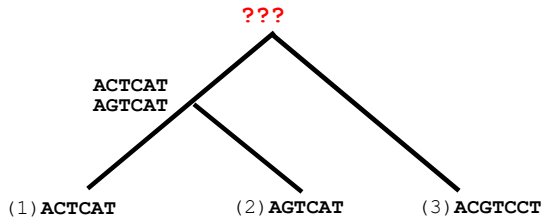
d(x,y) = 3
d(x,"_") = 2

(1) ACTCAT
(2) AGTCAT — 3

(2) A_GTCAT
(3) ACGTCCT — 5

(1) AC_TCAT
(3) ACGTCCT — 5

Progressive Alignment



- Use *profile alignment* to merge sequences according to a guide tree.
- Typically, the most similar sequence pairs are merged first.

Merging strategy:
Align the profile (1,2) with sequence (3)

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

d(x,y) = 3
d(x,"_") = 2

(1) ACTCAT
(2) AGTCAT — 3

(2) A_GTCAT
(3) ACGTCCT — 5

(1) AC_TCAT
(3) ACGTCCT — 5

```
  _ A C T C A T
  _ A G T C A T
    4
```

d(x,y) = 3
d(x,"_") = 2

```
_
A
C
G
T
C
C
T
```

A
A
_

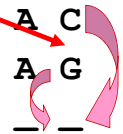
```
  _ A C T C A T
  _ A G T C A T
    4   8
```

d(x,y) = 3
d(x,"_") = 2

```
_
A
C
G
T
C
C
T
```

A C
A G
_ _

Note: no penalty for substitutions in the profile. We paid for those in a previous step



4+4

		A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8				
_							
A							
C							
G							
T							
C							
C							
T							

d(x,y) = 3
d(x,"_") = 2

		A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4						
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

		A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4	0					
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4	0	4				
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4	0	4	8			
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4	0	4	8	12		
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
		4	8	12	16	20	24
_							
A	4	0	4	8	12	16	
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
			4	8	12	16	20	24
_								
A	4	0	4	8	12	16	20	
C	8	4						
G	12							
T	16							
C	...							
C								
T								

A _
A _
A C

d(x,y) = 3
d(x,"_") = 2

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
			4	8	12	16	20	24
_								
A	4	0	4	8	12	16	20	
C	8	4	3					
G	12							
T	16							
C	...							
C								
T								

A C
A G
A C

d(x,y) = 3
d(x,"_") = 2

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
			4	8	12	16	20	24
_								
A	4	**0**	4	8	12	16	20	
C	8	**4**	**3**	7	8	12	16	
G	12	8	**7**	9	12	14	18	
T	16	12	11	**7**	11	15	14	
C	20	16	15	11	**7**	11	15	
C	24	20	19	15	11	**13**	17	
T	28	24	23	19	15	17	**13**	

d(x,y) = 3
d(x,"_") = 2

Two solutions:

ACGTCCT
A_CTCAT
A_GTCAT

ACGTCCT
AC_TCAT
AG_TCAT

Result:

Progressive alignment of the profile (1,2) with sequence (3)

(1,2) + (3)

(1) ACTCAT
(2) AGTCAT

(3) ACGTCCT

(1) AC_TCAT
(2) AG_TCAT
(3) ACGTCCT

4m+2g = 16

d(x,y) = 3
d(x,"_") = 2

		_	A	C	T	C	A	T
		_	A	_	T	C	A	T
		4	6	10	14	18	22	
_								
A	4	0						
C	8							
G	12							
T	16							
C	...							
C								
T								

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

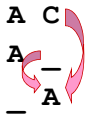
		_	A	C	T	C	A	T
		_	A	_	T	C	A	T
		4	6	10	14	18	22	
_								
A	4	0	2					
C	8							
G	12							
T	16							
C	...							
C								
T								

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

		_	A	C	T	C	A	T
		_	A	_	T	C	A	T
		4	6	10	14	18	22	
_								
A	4	0						
C	8							
G	12							
T	16							
C	...							
C								
T								

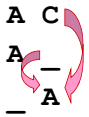
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

		_	A	C	T	C	A	T
		_	A	_	T	C	A	T
		4	6	10	14	18	22	
_								
A	4	0	9					
C	8							
G	12							
T	16							
C	...							
C								
T								

d(x,y) = 3
d(x,"_") = 2

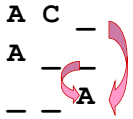


How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0					
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3

d(x,"_") = 2

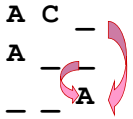


How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	10				
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3

d(x,"_") = 2

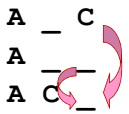


How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3

d(x,"_") = 2

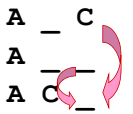


How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4	6				
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3

d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

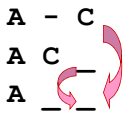
	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	4	8	12	16	20
C	8	4	2				
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

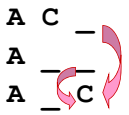
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	8	12	16	20	24
A	4	0	2	6	10	14	18
C	8	4	6				
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2



How would you score this case?

Progressive alignment

- "Once a gap, always a gap"
  - You can't go back and correct a bad decision at an earlier step.
- Progressive alignment is not guaranteed to give the optimal alignment.
- But it does have better complexity…

Optimal Pairwise Alignments

(1) ACTCAT
(2) AGTCAT

(2) A_GTCAT
(3) ACGTCCT

(1) AC_TCAT
(3) ACGTCCT

Progressive alignment

(1,2) + (3)

(3) ACGTCCT
(1) AC_TCAT
(2) AG_TCAT

4m+2g = 16

An alternate alignment

(1) AC_TCAT
(2) A_GTCAT
(3) ACGTCCT

2m+4g = 14

Scoring when the profile contains gaps

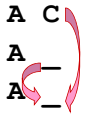
	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
		4	6	10	14	18	22
_							
A	4	0					
C	8						
G	12						
T	16						
C	…						
C							
T							

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
		4	6	10	14	18	22
_							
A	4	0	2				
C	8						
G	12						
T	16						
C	...						
C							
T							

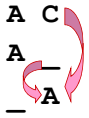
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
		4	6	10	14	18	22
_							
A	4	0					
C	8						
G	12						
T	16						
C	...						
C							
T							

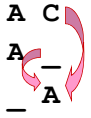
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
		4	6	10	14	18	22
_							
A	4	0	9				
C	8						
G	12						
T	16						
C	...						
C							
T							

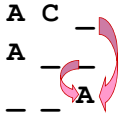
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
		4	6	10	14	18	22
_							
A	4	0					
C	8						
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2



How would you score this case?

		A	C	T	C	A	T
		4	6	10	14	18	22
A	4	0	10				
C	8						
G	12						
T	16						
C	...						
C							
T							

A C _
A _ _
_ _ A

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

		A	C	T	C	A	T
		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

A _ C
A _ _
A C _

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

		A	C	T	C	A	T
		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4	6				
G	12						
T	16						
C	...						
C							
T							

A _ C
A _ _
A C _

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

		A	C	T	C	A	T
		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

A C
A _
A C

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

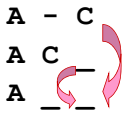
	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	4	8	12	16	20
C	8	4	2				
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	6	10	14	18	22
A	4	0	2	6	10	14	18
C	8	4					
G	12						
T	16						
C	...						
C							
T							

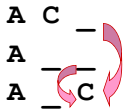
d(x,y) = 3
d(x,"_") = 2



How would you score this case?

	_	A	C	T	C	A	T
	_	A	_	T	C	A	T
_		4	8	12	16	20	24
A	4	0	2	6	10	14	18
C	8	4	6				
G	12						
T	16						
C	...						
C							
T							

d(x,y) = 3
d(x,"_") = 2

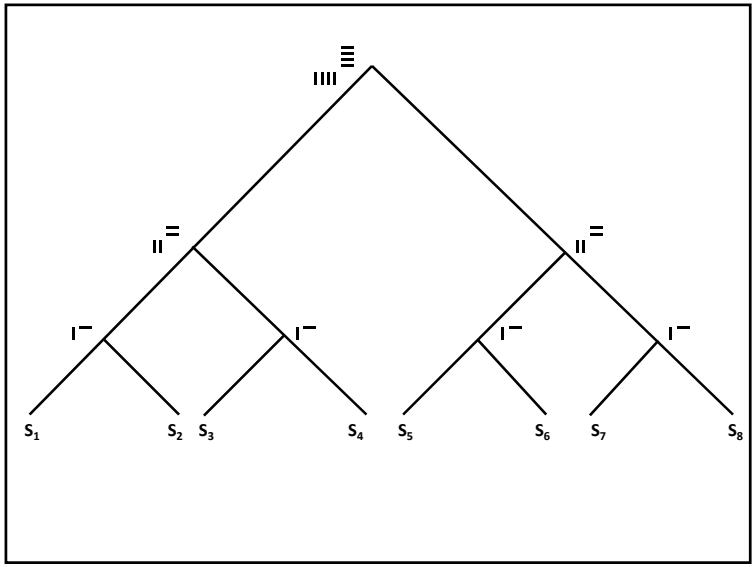


How would you score this case?

Complexity of progressive alignment

- Distance matrix
  - Each pairwise alignment $O(n^2)$
  - Number of pairwise alignments $O(k^2)$
- Iterative construction of MSA
  - Merging by profile alignment $O(k^2n^2)$
    (see class notes for details)

Entire method $O(k^2n^2)$



Summary: Progressive alignment heuristics

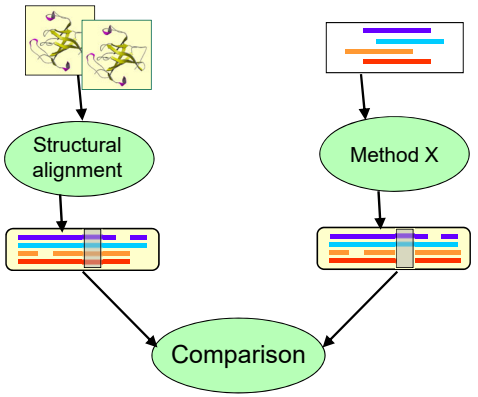
- Not guaranteed to give the optimal MSA
- Bad choice of gaps propagates
- Complexity
  - Progressive: $O(k^2n^2)$
  - versus DP: $O(n^k 2^k k^2)$
- Typically, merge the most closely related sequences first.

Mathematical correctness is not a guarantee of biological accuracy.

The performance of MSA programs is typically evaluated using benchmarks based on biological data:

- Curated structural alignment
- Automated structural alignment
- Real or simulated sequence

Multiple sequence alignment methods are evaluated against trusted benchmarks



Note that implementation choices result in substantial differences in running time:

Aligner	Performance*	Time
DIALIGN	57.2	12 h, 25 min
CLUSTALW	58.9	2 h, 57 min
T-Coffee	63.6	***144 h, 51 min***
MUSCLE	64.8	3 h, 11 min
MAFFT	64.8	***2h,36min***
ProbCons	66.9	19 h, 41 min
ProbCons-ext	***68.0***	37 h, 46 min

* Fraction of correctly aligned residue pairs

Do et al, Genome Research, 2005

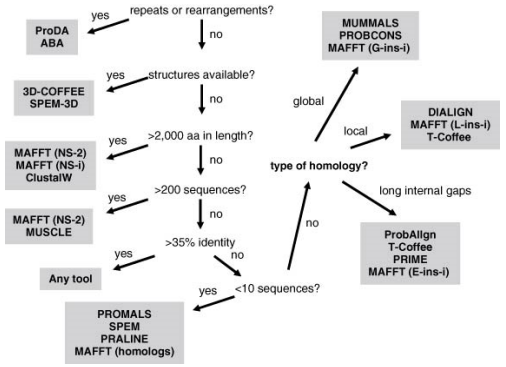
Table 9. Summary of the benchmarking results of several MSA program on BAliBASE 3.0 [123]

Program	R1-1	R1-2	R2	R3	R4	R5	Avg Score	Tot time(s)
MSAProbs [122]	0.441	**0.865**	**0.464**	**0.607**	**0.622**	**0.608**	**0.607**	12382.00
Probalign [84]	**0.453**	0.862	0.439	0.566	0.603	0.549	0.589	10095.20
MAFFT [14]	0.439	0.831	0.450	0.581	0.605	0.591	0.588	1475.40
Probcons [17]	0.417	0.855	0.406	0.544	0.532	0.573	0.558	13086.30
Clustal Omega [123]	0.358	0.789	0.450	0.575	0.579	0.533	0.554	539.91
T-Coffee [71]	0.410	0.848	0.402	0.491	0.545	0.587	0.551	81041.50
Kalign [72]	0.365	0.790	0.360	0.476	0.504	0.435	0.501	**21.88**
MUSCLE [15]	0.318	0.804	0.350	0.409	0.450	0.460	0.475	789.57
FSA [136]	0.270	0.818	0.187	0.259	0.474	0.398	0.419	53648.10
DiAlign [12]	0.265	0.696	0.292	0.312	0.441	0.425	0.415	3977.44
PRANK [137]	0.223	0.680	0.257	0.321	0.360	0.356	0.376	128355.00
ClustalW [11]	0.227	0.712	0.220	0.272	0.396	0.308	0.374	766.47

Several benchmarking results on BAliBASE 3.0 are summarized. This table presents total column (TC) scores for six references, average TC scores on all references, and total running times. The best results in each column are put in bold.

Zhang *et al*, Briefings in Bioinformatics, 2022, 23(3), 1–162022

Which program to choose?



Do and Katoh, 2008