Multiple alignment: ClustalW


- Pairwise alignment (fast approximation or full dynamic programming)
- Generate a distance matrix (% identities converted to distances)
- Construct a guide tree (neighbor-joining phylogenetic tree)
- Progressive alignment following the guide tree (scoring matrix, sequence weight, gap penalties, etc.)

Multiple alignment: ClustalW

- Generate a distance matrix from % identities
- Sequence weights: based on branch lengths

…

Multiple alignment: ClustalW

- Rooted at the midpoint
- Shorter branch lengths → Fewer changes (e.g., S1 and S2 are more similar than S1 and S3)
Multiple alignment: ClustalW

- Sequence weights: based on branch lengths

*Weights are normalized so that the max becomes 1.0

Root

Divided by the number of sequences sharing the branch

- Shorter branch lengths
  - Fewer changes

Closely related sequences get smaller weights

These sequences share duplicated information

Multiple alignment: ClustalW

- Progressive alignment following the guide tree

Profile alignment: alignment vs. sequence, alignment vs. alignment

Profile scores:

\[ \text{Score} = \frac{\text{W1} \times \text{W3} \times \text{S}(P,E) + \text{W2} \times \text{W3} \times \text{S}(G,E)}{2} \]

[Without weighting]

\[ \text{Score} = \frac{\text{S}(P,E) + \text{S}(G,E)}{2} \]

[With weighting]

\[ \text{W1} = 0.02 + \frac{0.15}{2} + \frac{0.09}{3} = 0.125 \]
\[ \text{W2} = 0.08 + \frac{0.15}{2} + \frac{0.09}{3} = 0.185 \]
\[ \text{W3} = 0.38 + \frac{0.09}{3} = 0.410 \]
\[ \text{W4} = 0.460 \]
Multiple alignment: ClustalW

How sequence weighting works: Example 1

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>V</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>V</td>
<td>A</td>
</tr>
</tbody>
</table>

Score = \{S(V,P) + S(V,P)\} / 2

Without weighting:

\[\text{Score} = \frac{3 + 3}{2} = 3\]

With weighting:

\[\text{Score} = \frac{1.1 \times 3 + 1.1 \times 3}{2} = 3.3\]

How sequence weighting works: Example 2

How should 'VG' in S4 be aligned?

S1 VAG or VAG or VGA
S2 VGA or VGA or VGA
S3 IVG or IVG or IVG
S4 VG- or V-G or V-G

Position 1 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = gap penalty x 3
(c) Score = {1.1x3} / 3

Position 2 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 3 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 4 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Multiple alignment: ClustalW

How should 'VG' in S4 be aligned?

S1 VAG or VAG or VGA
S2 VGA or VGA or VGA
S3 IVG or IVG or IVG
S4 VG- or V-G or V-G

Position 1 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = gap penalty x 3
(c) Score = {1.1x3} / 3

Position 2 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 3 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 4 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Multiple alignment: ClustalW

How should 'VG' in S4 be aligned?

S1 VAG or VAG or VGA
S2 VGA or VGA or VGA
S3 IVG or IVG or IVG
S4 VG- or V-G or V-G

Position 1 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = gap penalty x 3
(c) Score = {1.1x3} / 3

Position 2 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 3 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 4 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Multiple alignment: ClustalW

How should 'VG' in S4 be aligned?

S1 VAG or VAG or VGA
S2 VGA or VGA or VGA
S3 IVG or IVG or IVG
S4 VG- or V-G or V-G

Position 1 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = gap penalty x 3
(c) Score = {1.1x3} / 3

Position 2 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 3 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3

Position 4 (match=2, mismatch=1)

(a) Score = {S(V,P)} / 2
(b) Score = {1.1x3} / 3
(c) Score = gap penalty x 3
Progressive alignment following the guide tree
• Which scoring matrix to use?
  → Determined based on distance between sequences
    80 - 100% identity → Blosum80
    60 - 80% identity → Blosum62
    30 - 60% identity → Blosum45
    0 - 30% identity → Blosum30

Multiple alignment: ClustalW
[Progressive alignment following the guide tree]
• How gap penalties are determined?
  → Initial gap penalties: GOP (gap opening) and GEP (gap extension) → set by the user
  → Weight (scoring) matrix dependent gap penalties
  → Similarity level dependent gap penalties
  → Sequence length dependent gap penalties
  → Position specific gap penalties
  • if gaps already exist
  • residue specific (e.g., hydrophilic stretches)


Profile alignment
• Progressive alignment
  → Greedy (finds local optima, but no guarantee for global optima)
  → Errors (incorrect gap positions) in the early alignments cannot be rectified later
• Global alignment only (local similarity may be missed)

Multiple alignment: ClustalW
[Progressive alignment following the guide tree]

Multiple alignment: ClustalW Web servers
[Legacy version: use Clustal-Omega instead]
https://galaxy.pasteur.fr/ (part of Galaxy @Pasteur)
http://www.ebi.ac.uk/software/ClustalW.html

Clustal-Omega:
http://www.ebi.ac.uk/Tools/msa/clustalw/
http://www.clustal.org/ (Clustal original website)

Multiple alignment: ClustalX
ClustalX FTP site (Windows, Macintosh, Linux/Unix)
http://www.clustal.org/
• ClustalW2 with GUI; can be used independently
  → includes alignment quality analysis functions
  → can be used as an alignment format converter

Multiple alignment: Heuristic methods

“Multiple Sequence Alignment Methods” (2014) in Methods in Molecular Biology (Russell, ed.), vol. 1079
[Downloadable from Springer Link, Canvas]
Multiple alignment: T-Coffee

T-Coffee Home page
http://tcoffee.crg.cat/
http://www.tcoffee.org
http://tcoffee.vital-it.ch/apps/tcoffee/index.html
(mirror site)
http://www.ebi.ac.uk/Tools/msa/tcoffee/


Primary libraries of alignments

- Consistent global and local alignments are stacked.
- Weight of the stacked alignment is sum of the weights of original alignments.
- Each residue pair in an alignment is weighted based on the alignment weight.
- The weight of consistent alignments in the library are also added for the residue pair found in the aligned region.

Extended library

- Pre-made scoring matrices (BLOSUM etc.) are not used.
- No gap penalty is used.
Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

Global pairwise alignment
Primary Library

Local pairwise alignment
Primary Library

Consistent global and local alignments are stacked.

Pooled
Primary Library

Extended Library

Weight of the stacked alignment = sum of the weights of original alignments

T-Coffee does not use pre-made scoring matrices (BLOSUM etc.)
or gap penalties

Progressive Alignment with NJ guide tree
using the generated scoring matrix
for each pair of sequences

http://tcoffee.crg.cat/apps/tcoffee/all.html

Multiples alignment: T-Coffee

http://tcoffee.crg.cat/apps/tcoffee/all.html

T-Coffee: Aligns DNA, RNA, or protein sequences

R-Coffee: Aligns RNA sequences considering their 2D/3D structures

Expresso (3D-Coffee): Aligns protein sequences with structural information incorporated as pairwise structural alignments (library)

PSI-Coffee: Extends each sequence with PSI-BLAST
- Structural information can be incorporated, too
- PSI/TM-Coffee incorporates TM prediction, too

Pre-Coffee: Promoter region alignment
**Multiple alignment: T-Coffee**

* http://tcoffee.crg.cat/apps/tcoffee/all.html

**M-Coffee:** Combines multiple alignments
- Shows consistency between multiple alignments

**Combine:** Combines your own multiple alignments

---

**Evaluating multiple alignments using TCS**


**T-Coffee**

**ClustalW2**

---

**T-Coffee**

- M-Coffee: Combines multiple alignments
- Shows consistency between multiple alignments

**Combine:** Combines your own multiple alignments

---

**Evaluating multiple alignments using TCS**


**T-Coffee**

**ClustalW2**

---

**T-Coffee**

- M-Coffee: Combines multiple alignments
- Shows consistency between multiple alignments

**Combine:** Combines your own multiple alignments

---

**Evaluating multiple alignments using TCS**


**T-Coffee**

**ClustalW2**

---

**T-Coffee**

- M-Coffee: Combines multiple alignments
- Shows consistency between multiple alignments

**Combine:** Combines your own multiple alignments

---

**Evaluating multiple alignments using TCS**


**T-Coffee**

**ClustalW2**

---