Multiple Sequence Alignment

Ian Chen

University of Illinois Urbana-Champaign

Examples

Sequence Similarity

Which pair of sequences are the closest?

$$S_1 = AAAAA \qquad S_2 = AGAGA \qquad S_3 = GGAAA$$

Examples

Sequence Similarity

Which pair of sequences are the closest?

$$S_1 = AAAAA \qquad S_2 = AGAGA \qquad S_3 = GGAAA$$

Examples

Orthology Detection

$$\begin{pmatrix} S_1 & L & A & S & T & F & A & - & T & C & A & T \\ S_2 & L & A & S & T & C & A & - & T & - & - & - \\ S_3 & V & E & R & Y & F & A & S & T & C & A & T \\ S_4 & - & - & - & - & F & A & - & T & C & A & T \end{pmatrix}$$

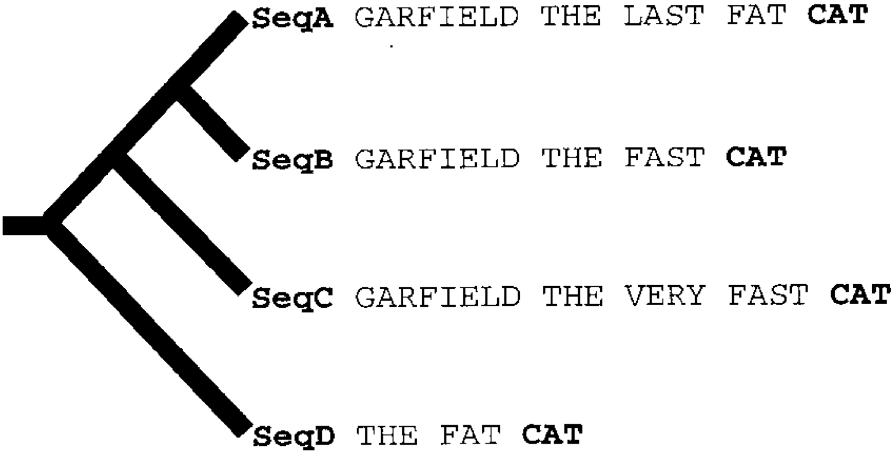
Examples

Orthology Detection

$$\left(\begin{array}{c|cccccccccccc} S_1 & L & A & S & T & F & A & - & T & C & A & T \\ S_1 & L & A & S & T & - & - & - & - & C & A & T \\ S_3 & V & E & R & Y & F & A & S & T & C & A & T \\ S_4 & - & - & - & - & F & A & - & T & C & A & T \end{array}\right)$$

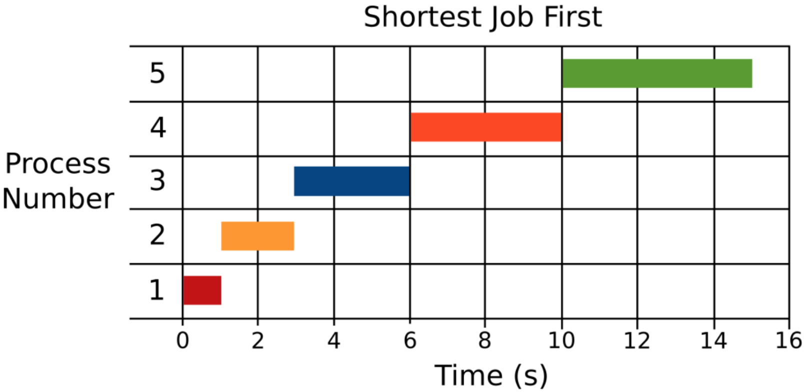
Examples

Phylogeny Estimation



Examples

Scheduling



Outline

Multiple Sequence Alignment

Pairwise Alignments

Alignment Graphs

Maximum Weight Trace

T-COFFEE

MAGUS

Results

Questions

Multiple Sequence Alignment

Problem (Multiple Sequence Alignment)

Given a set of sequences $S_1, \ldots, S_n$, and a scoring function d, find a *multiple sequence alignment* $\mathscr{A} = (a_{ij})_{1 \leq i \leq n}$ minimizing

$$\sum_j d(a_{1j}, \ldots, a_{nj})$$

Multiple Sequence Alignment

Problem (Multiple Sequence Alignment)

Given a set of sequences $S_1, \ldots, S_n$, and a scoring function d, find a *multiple sequence alignment* $\mathscr{A} = (a_{ij})_{1 \leq i \leq n}$ minimizing

$$\sum_j d(a_{1j}, \ldots, a_{nj})$$

$$d(\cdot) = \begin{cases} +\infty & \text{if number unique characters is more than 1} \\ 1 & \text{otherwise} \end{cases}$$

Multiple Sequence Alignment

Problem (Multiple Sequence Alignment)

Given a set of sequences $S_1, \ldots, S_n$, and a scoring function d, find a *multiple sequence alignment* $\mathscr{A} = (a_{ij})_{1 \leq i \leq n}$ minimizing

$$\sum_j d(a_{1j}, \ldots, a_{nj})$$

$$d(\cdot) = \begin{cases} 0 & \text{if number unique characters is more than 1} \\ -1 & \text{otherwise} \end{cases}$$

Multiple Sequence Alignment

Problem (Multiple Sequence Alignment)

Given a set of sequences $S_1, \ldots, S_n$, and a scoring function d, find a *multiple sequence alignment* $\mathscr{A} = (a_{ij})_{1 \leq i \leq n}$ minimizing

$$\sum_j d(a_{1j}, \ldots, a_{nj})$$

$$d(\cdot) = \# \text{ dashes} + \# \text{ unique} - 1$$

Pariwise Alignments

Edit Distance

	λ	T	T	A	A	G	C
λ	**0**	1	2	3	4	5	6
A	**1**	1	2	2	3	4	5
A	**2**	2	2	2	2	3	4
T	3	**2**	2	3	3	3	4
T	4	3	**2**	2	3	4	4
A	5	4	3	**2**	2	3	4
A	6	5	4	3	**2**	3	4
G	7	6	5	4	3	**2**	**3**

Alignment Graphs

Definition (Alignment Graph)

Given a set of sequences $S_1, \ldots, S_n$, and a scoring function d, and a set of *pairwise* alignments $A_1, \ldots, A_k$, construct $G = (V, E, \prec)$, where

1. For each sequence, for each *site* s_{ij}, create a vertex
2. For each alignment A_i, for each *homology* s_{ij}, s_{kl}, add weight $d(s_{ij}, s_{kl})$ to the edge
3. For each pair site s_{ij} and $s_{ij'}$ where $j' > j$, add $s_{ij}, s_{ij'}$ to $\prec$

Alignment Graphs

$$S_1 = ABCD \qquad S_2 = BAC \qquad S_3 = AAD$$

$$A_1 = \left(\begin{array}{c|ccccc} S_1 & A & B & - & C & D \\ S_2 & - & B & A & C & - \end{array}\right)$$

$$A_2 = \left(\begin{array}{c|cccc} S_2 & B & A & - & C \\ S_3 & - & A & A & D \end{array}\right)$$

$$A_3 = \left(\begin{array}{c|ccccc} S_1 & - & A & B & C & D \\ S_3 & A & A & - & - & D \end{array}\right)$$

Alignment Graphs

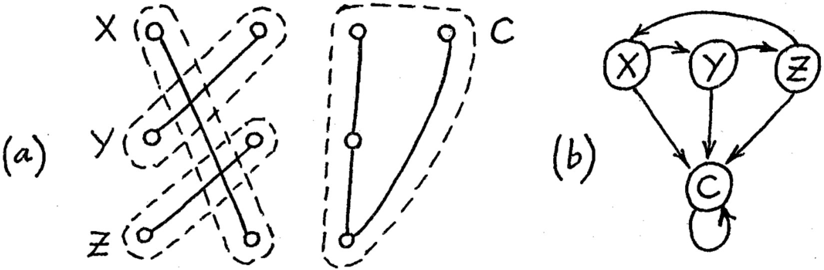


Fig. 1. (a) An alignment graph on three sequences. We use the convention of drawing the characters in a sequence horizontally left to right. (b) Relation $\prec^*$ on its connected components.

Alignment Graphs

Definition (Trace)

A *trace* of an alignment graph $G = (V, E, \prec)$ is a subset of the $T \subset E$ where $G^* = (V, T, \prec^*)$ is acyclic.

$$\mathscr{A} = \left(\begin{array}{c|ccccc} S_1 & A & B & C & - & D \\ S_2 & - & B & A & C & - \\ S_3 & - & - & A & A & D \end{array}\right)$$

Maximum Weight Trace

Problem (Maximum Weight Trace (MWT))

Given an alignment graph $G = (V, E, \prec)$, find the trace T that *maximizes*

$$\sum_{e \in T} w(e)$$

Maximum Weight Trace

Theorem (Kececioglu'93)

Maximum Weight Trace is NP-Hard

Proof.

Consider an instance $G = (V, E)$ and integer k of *Feedback Set*.

1. For every vertex v, create sequence $S_v = v$
2. For every edge $u \to v$, create sequence $S_{uv} = uv$
3. Create pairwise alignments

$$\left(\begin{array}{c|cc} S_u & u & - \\ S_{uv} & u & v \end{array}\right) \qquad \left(\begin{array}{c|cc} S_v & - & v \\ S_{uv} & u & v \end{array}\right)$$

4. Check if MWT is at least $2k$.

□

Maximum Weight Trace

Theorem (Kececioglu'93)

Maximum Weight Trace is NP-Hard

Proof.

Consider an instance $G = (V, E)$ and integer k of *Feedback Set*.

1. For every vertex v, create sequence $S_v = v$
2. For every edge $u \to v$, create sequence $S_{uv} = uv$
3. Create pairwise alignments

$$\begin{pmatrix} S_u & u & - \\ S_{uv} & u & v \end{pmatrix} \qquad \begin{pmatrix} S_v & - & v \\ S_{uv} & u & v \end{pmatrix}$$

4. Check if MWT is at least $2k$.

□

Maximum Weight Trace

Theorem (Kececioglu'93)

Maximum Weight Trace is NP-Hard

Proof.

Consider an instance $G = (V, E)$ and integer k of *Feedback Set*.

1. For every vertex v, create sequence $S_v = v$
2. For every edge $u \to v$, create sequence $S_{uv} = uv$
3. Create pairwise alignments

$$\begin{pmatrix} S_u & u & - \\ S_{uv} & u & v \end{pmatrix} \qquad \begin{pmatrix} S_v & - & v \\ S_{uv} & u & v \end{pmatrix}$$

4. Check if MWT is at least $2k$.

□

Maximum Weight Trace

Theorem (Kececioglu'93)

Maximum Weight Trace is NP-Hard

Proof.

Consider an instance $G = (V, E)$ and integer k of *Feedback Set*.

1. For every vertex v, create sequence $S_v = v$
2. For every edge $u \to v$, create sequence $S_{uv} = uv$
3. Create pairwise alignments

$$\begin{pmatrix} S_u & u & - \\ S_{uv} & u & v \end{pmatrix} \qquad \begin{pmatrix} S_v & - & v \\ S_{uv} & u & v \end{pmatrix}$$

4. Check if MWT is at least $2k$.

□

Maximum Weight Trace

Let $D(x_1, \ldots, x_n)$ denote maximum weight trace over *prefixes* $S_i[1 : x_i]$. Then,

$$D(\vec{x}) = \max_{\vec{b} \in [2]^n} \{ D(\vec{x} - \vec{b}) + d(\vec{S}^{\vec{b}}) \}$$

Thus, MWT can be solved in $O((2k)^n poly(n))$.

Using the *Branch-and-Bound* paradigm, this can be fast.

Maximum Weight Trace

Let $D(x_1, \ldots, x_n)$ denote maximum weight trace over *prefixes* $S_i[1 : x_i]$. Then,

$$D(\vec{x}) = \max_{\vec{b} \in [2]^n} \{ D(\vec{x} - \vec{b}) + d(\vec{S}^{\vec{b}}) \}$$

Thus, MWT can be solved in $O((2k)^n poly(n))$.

From Fa'24 CS 374 homework 14, this can be improved to $O(nk^n poly(n))$.

Using the *Branch-and-Bound* paradigm, this can be fast.

Σ

Maximum Weight Trace

Let $D(x_1, \ldots, x_n)$ denote maximum weight trace over *prefixes* $S_i[1 : x_i]$. Then,

$$D(\vec{x}) = \max_{\vec{b} \in [2]^n} \{ D(\vec{x} - \vec{b}) + d(\vec{S}^{\vec{b}}) \}$$

Thus, MWT can be solved in $O((2k)^n poly(n))$.

Using the *Branch-and-Bound* paradigm, this can be fast.

T-COFFEE

1. Aggregate pairwise alignments through triples
2. Calculate new pairwise distances
3. Compute guide trees
4. Progressively align using guide tree

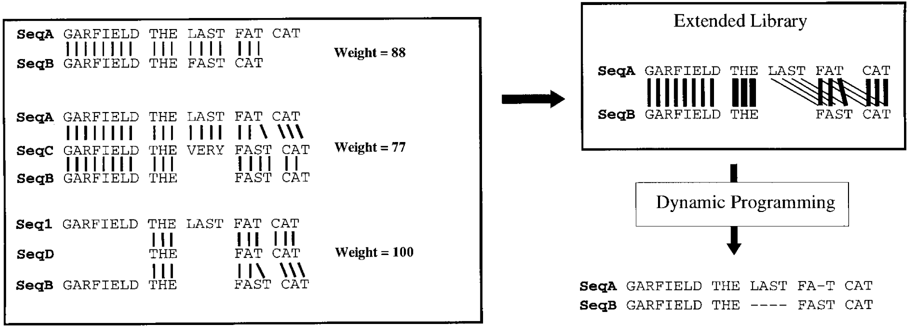
T-COFFEE

b)Primary Library

```
SeqA GARFIELD THE LAST FAT CAT   Prim. Weight = 88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT  Prim. Weight = 77
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT   Prim. Weight =100
SeqD -------- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT  Prim Weight = 100
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT       Prim. Weight = 100
SeqD -------- THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT  Prim. Weight = 100
SeqD -------- THE ---- FA-T CAT
```

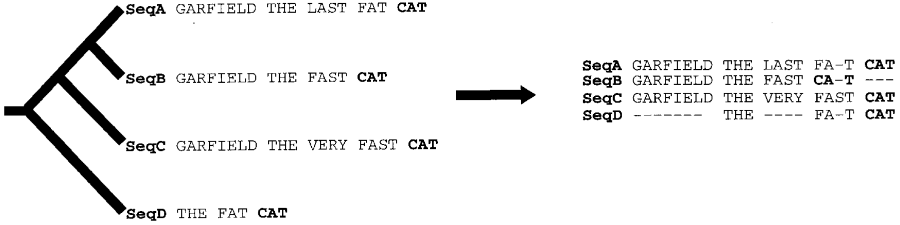
T-COFFEE

c)Extended Library for seq1 and seq2



T-COFFEE

a)Regular Progressive Alignment Strategy



MAGUS

1. Create alignment graph from backbone alignments
2. Cluster with Markov Clustering (MCL)
3. Break all clusters that violate ordering

MAGUS

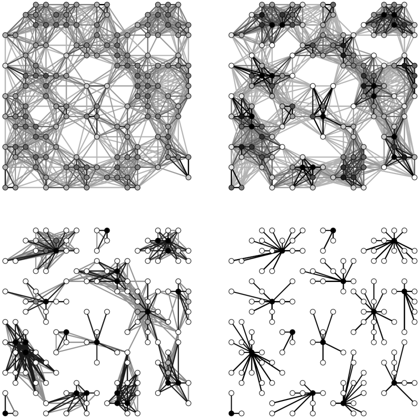


Figure 3. Successive stages of flow simulation by the *MCL* process.

Clusters have high *edge connectivity*. A random walk is *likely* to stay within the cluster.

Markov Clustering Algorithm

1. Expansion (random walk)
2. Inflation (amplify probabilities)
3. Repeat

MAGUS

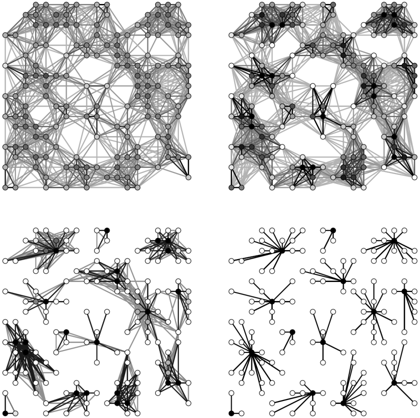


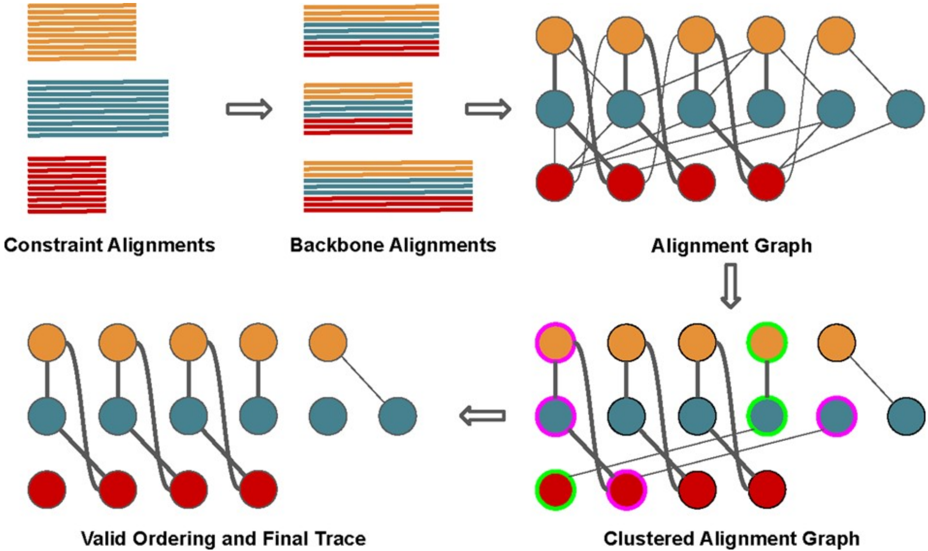
Figure 3. Successive stages of flow simulation by the *MCL* process.

Clusters have high *edge connectivity*. A random walk is *likely* to stay within the cluster.

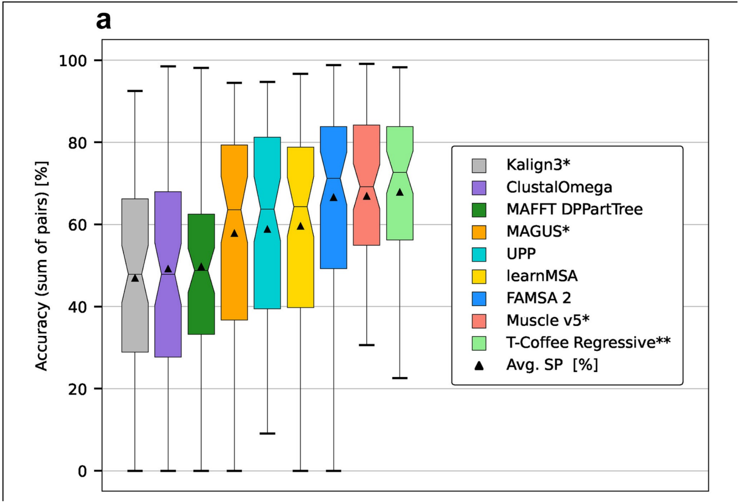
Markov Clustering Algorithm

1. Expansion (random walk)
2. Inflation (amplify probabilities)
3. Repeat

MAGUS


Constraint Alignments
Backbone Alignments
Alignment Graph
Clustered Alignment Graph
Valid Ordering and Final Trace

Results


a
Accuracy (sum of pairs) [%]
100
80
60
40
20
0
Kalign3*
ClustalOmega
MAFFT DPPartTree
MAGUS*
UPP
learnMSA
FAMSA 2
Muscle v5*
T-Coffee Regressive**
Avg. SP [%]

Results

1. T-Coffee (Regressive) is the best
2. Consistency based methods are very good (MAGUS, T-Coffee, Muscle)
3. Single-stage aligners are bad (Kalign, ClustalOmega, ...)
4. Exception for FAMSA

Questions

1. How can we encode genome events into the alignment graph?
2. Can T-COFFEE perform better if we give it multiple sequence alignments (instead of pairwise) as input?
3. Do other clustering algorithms (beyond MCL) cluster the alignment graph better in MAGUS?

Bibliography I

John Kececioglu.

The maximum weight trace problem in multiple sequence alignment.

In Alberto Apostolico, Maxime Crochemore, Zvi Galil, and Udi Manber, editors, *Combinatorial Pattern Matching*, pages 106–119, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa.

T-coffee: a novel method for fast and accurate multiple sequence alignment.

Journal of Molecular Biology, 302(1):205–217, 2000.

Bibliography II

Luisa Santus, Edgar Garriga, Sebastian Deorowicz, Adam Gudyś, and Cedric Notredame.

Towards the accurate alignment of over a million protein sequences: Current state of the art.

Current Opinion in Structural Biology, 80:102577, 2023.

Vladimir Smirnov and Tandy Warnow.

Magus: Multiple sequence alignment using graph clustering.

Bioinformatics, 37(12):1666–1672, 11 2020.

Bibliography III

Stijn Van Dongen.

Graph clustering via a discrete uncoupling process.

SIAM Journal on Matrix Analysis and Applications, 30(1):121–141, 2008.

Paul Zaharias, Vladimir Smirnov, and Tandy Warnow.

The maximum weight trace alignment merging problem.

In Carlos Martín-Vide, Miguel A. Vega-Rodríguez, and Travis Wheeler, editors, *Algorithms for Computational Biology*, pages 159–171, Cham, 2021. Springer International Publishing.