10 Progressive alignment

10.1 Introduction to progressive alignment

Several heuristic solutions to compute MSAs have been developed to avoid the multi-dimensional DP approach that requires heavy computational power.

Three cases of aligning multiple sequences

- Two sequences, e.g. s^1 and s^2
- One alignment and one sequence, e.g. $\mathcal{A}^1$ and s^1
- Two alignments, e.g. $\mathcal{A}^1$ and $\mathcal{A}^2$

Guiding methods

- Clustering
- Phylogenetic tree

Aligning methods

- Complete alignment
- Pair-guided alignment

Once a gap always a gap

Many progressive alignment procedures use the once a gap always a gap policy, hence it is difficult to fix the errors that are made in early steps.

10.2 Alignment clustering

Alignment clustering can be used even when accurate phylogenic trees are not available.

Clustering methods

- Linear clustering
- Linkage clustering

Linear clustering

1. Start with an alignment with a single sequence
2. Add a single sequence to the alignment
3. Repeat until no sequence is left

Selection of the next sequence

- Most similar to the one already in the alignment
- Most similar to the average sequence in the alignment

Pseudo-code of linear progressive alignment (general progressive alignment)

Algorithm 10.1: General progressive alignment

```
U: Set of sequences not aligned
$\mathcal{A}$: Current alignment

U ← $\{s_1, s_2, ...s_n\}$;
Choose two sequences s and t from U;
U ← U − $\{s, t\}$;
$\mathcal{A}$ ← $Align(s, t)$;

for $i$ ← 1 to $n - 2$ do
    Choose a sequence s from U;
    U ← U − $\{s\}$;
    $\mathcal{A}$ ← $Align(\mathcal{A}, s)$;
end
```

Linkage methods

It requires the pair-wise alignment scores of all possible combinations.

- Average linkage
- Maximum linkage
- Minimum linkage

Example of linkage methods

It requires the pair-wise alignment scores of all possible combinations.

Decide two alignments from the three alignments, $\mathcal{A}^1 = \{s^1\}, \mathcal{A}^2 = \{s^2\}$, and $\mathcal{A}^3 = \{s^3, s^4\}$, for clustering.

Pair-wise scores

	s1	s2	s3	s4
s1	0	7	5	3
s2		0	4	8
s3			0	2
s4				0

Linkage selection

$$\begin{array}{lll}
\text{Average linkage} & S(\mathcal{A}_1,\mathcal{A}_2)=7 & \checkmark \\
 & S(\mathcal{A}_1,\mathcal{A}_3)=(5+3)/2=4 & \\
 & S(\mathcal{A}_2,\mathcal{A}_3)=(4+8)/2=6 & \\
 & & \\
\text{Maximum linkage} & S(\mathcal{A}_1,\mathcal{A}_2)=7 & \\
 & S(\mathcal{A}_1,\mathcal{A}_3)=\max(5,3)=5 & \\
 & S(\mathcal{A}_2,\mathcal{A}_3)=\max(4,8)=8 & \checkmark \\
 & & \\
\text{Minimum linkage} & S(\mathcal{A}_1,\mathcal{A}_2)=7 & \checkmark \\
 & S(\mathcal{A}_1,\mathcal{A}_3)=\min(5,3)=3 & \\
 & S(\mathcal{A}_2,\mathcal{A}_3)=\min(4,8)=4 &
\end{array}$$

Exercise 10.1

Select two alignments from the three alignments: $\mathcal{A}^1=\{s^1\}$, $\mathcal{A}^2=\{s^2\}$, and $\mathcal{A}^3=\{s^3,s^4\}$ for clustering.

	s1	s2	s3	s4
s1	0	2	2	5
s2		0	4	5
s3			0	1
s4				0

1. Use the average linkage.

2. Use the maximum linkage.

3. Use the minimum linkage.

10.3 Aligning methods

The progressive alignment method keeps combining two alignments until it produces the final alignment.

Aligning methods for progressive alignment

- Complete alignment
- Pair-guided alignment
- Conesus alignment
- Profile alignment

Complete alignment

It uses DP with a two-dimensional array to find gap positions between two alignments.

The score of a cell at column j and row i can be calculated as:

$$S(i,j) = \frac{1}{nm} \sum_{p \in \{p_1 \dots p_n\}} \sum_{q \in \{q_1 \dots q_m\}} R(\bar{s}_i^p, \bar{s}_j^q).$$

where n and m are the size of alignments, and $R(\cdot,\cdot)$ is a score function.

N.B. Notice $R(-,-)$ is always 0.

Example of complete alignment

Combine two alignments, $\mathcal{A}^p$ and $\mathcal{A}^q$ with a simple scoring scheme: Match: 1, Mismatch: -1, and Gap penalty: 1.

$\mathcal{A}^p$

s^{p1}: `GAT`
s^{p2}: `G-T`

$\mathcal{A}^q$

s^{q1}: `GT`
s^{q2}: `A-`
s^{q3}: `AT`

DP table

		s^{q1}	G	T
		s^{q2}	A	-
		s^{q3}	A	T
s^{p1}	s^{p2}	0		
G	G			
A	-			
T	T			

Initialization

$$S(0,1) = \frac{1}{6}(-1 \times 6) \qquad = -1$$

$$S(0,2) = -1 + \frac{1}{6}(-1 \times 4) \qquad = -1.67$$

$$S(1,0) = \frac{1}{6}(-1 \times 6) \qquad = -1$$

$$S(2,0) = -1 + \frac{1}{6}(-1 \times 3) \qquad = -1.5$$

$$S(3,0) = -1.5 + \frac{1}{6}(-1 \times 6) \qquad = -2.5$$

Cell update: $S(1,1)$

$$S(1,1)^{(1)} = -1 - 1 = -2$$
$$S(1,1)^{(2)} = -1 - 1 = -2$$
$$\begin{aligned} S(1,1)^{(3)} &= \frac{1}{2 \times 3}((R(G,G) + R(G,A) + R(G,A)) + (R(G,G) + R(G,A) + R(G,A))) \\ &= \frac{1}{6}((1 - 1 - 1) + (1 - 1 - 1)) = -0.33 \end{aligned}$$

DP table after $S(1,1)$ **update**

		s^{q1}	G	T
		s^{q2}	A	-
		s^{q3}	A	T
s^{p1}	s^{p2}	0	-1	-1.67
G	G	-1	-0.33	
A	-	-1.5		
T	T	-2.5		

Pair-guided alignment

Pair-guide alignment uses two sequences from two different alignments.

Example of pair-guided alignment

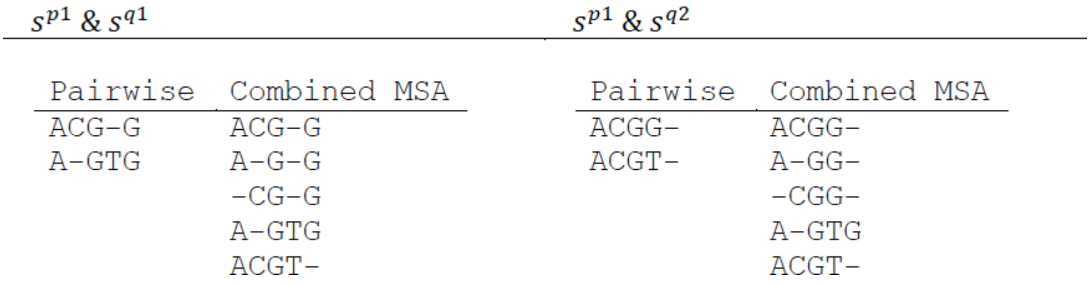
Combine two alignments, $\mathcal{A}^p$ and $\mathcal{A}^q$.

$\mathcal{A}^p$

s^{p1}:	ACGG
s^{p2}:	A-GG
s^{p3}:	-CGG

$\mathcal{A}^q$

s^{q1}:	A-GTG
s^{q2}:	ACGT-



s^{p1} & s^{q1}

Pairwise	Combined MSA
ACG-G	ACG-G
A-GTG	A-G-G
	-CG-G
	A-GTG
	ACGT-

s^{p1} & s^{q2}

Pairwise	Combined MSA
ACGG-	ACGG-
ACGT-	A-GG-
	-CGG-
	A-GTG
	ACGT-

Exercise 10.2

Combine two alignments $\mathcal{A}^p$ and $\mathcal{A}^q$ by using the pair-guided approach.

$\mathcal{A}^p$

s^{p1}: `TCG`
s^{p2}: `-CG`
s^{p3}: `T-C`

$\mathcal{A}^q$

s^{q1}: `T-G`
s^{q2}: `ACG`

1. Use the alignment between s^{p3} and s^{q2}.

s^{p3}: `T-C-`
s^{q2}: `-ACG`

10.4 CLUSTAL

CLUSTAL W is the most widely used progressive alignment program.

Original version (CLUSTAL)

- Pairwise alignment between all sequence pairs
- Phylogenic tree by UPGMA
- Guided by phylogenetic tree
- Align by consensus sequences

CLUSTAL W

- Phylogenic tree by Neighbor-joining
- Align by profiles

Gap penalty

- Open
- Extend
- End
- Separation

Web version

- `http://www.ch.embnet.org/software/ClustalW.html`