Multiple Sequence Alignment:
Hueristics; A 2–Approximation Algorithm

Prof. Benny Chor
School of Computer Science
Tel-Aviv University

Based in part on MSA sections in Gusfield’s book, and chapter 3 in Kanehisa’s book
MSA is Hard

• Input: Sequences $S_1, S_2, \ldots, S_k$, $k \geq 3$. 

• Typical sequences' lengths $n_1, n_2, \ldots, n_k \approx a$ few hundreds for AAs, 1000-2000 bp for DNA sequences.

• DP on the full $k$-dim box of volume $n_1 \times n_2 \times \ldots \times n_k$ takes $O(n_1 \cdot n_2 \cdot \ldots \cdot n_k \cdot 2^k)$.

• Such running time is very slow even for $k = 3$ and totally infeasible for $k \geq 6$.

• Certain versions of MSA are known to be NPH, so an exact poly time (poly in what?) is unlikely.
MSA is Hard

- Input: Sequences $S_1, S_2, \ldots, S_k$, $k \geq 3$.
- Typical sequences’ lengths $n_1, n_2, \ldots, n_k \approx$ a few hundreds for AAs, 1000-2000 bp for DNA sequences.
MSA is Hard

- Input: Sequences $S_1, S_2, \ldots, S_k$, $k \geq 3$.
- Typical sequences’ lengths $n_1, n_2, \ldots, n_k \approx$ a few hundreds for AAs, 1000-2000 bp for DNA sequences.
- DP on the full $k$-dim box of volume $n_1 \times n_2 \times \ldots \times n_k$ takes $O(n_1 \cdot n_2 \cdot \ldots \cdot n_k \cdot 2^k)$. 
MSA is Hard

- Input: Sequences $S_1, S_2, \ldots, S_k$, $k \geq 3$.
- Typical sequences’ lengths $n_1, n_2, \ldots, n_k \approx$ a few hundreds for AAs, 1000-2000 bp for DNA sequences.
- DP on the full $k$-dim box of volume $n_1 \times n_2 \times \ldots \times n_k$ takes $O(n_1 \cdot n_2 \cdot \ldots \cdot n_k \cdot 2^k)$.
- Such running time is very slow even for $k = 3$, and totally infeasible for $k \geq 6$. 
MSA is Hard

• Input: Sequences $S_1, S_2, \ldots, S_k$, $k \geq 3$.

• Typical sequences’ lengths $n_1, n_2, \ldots, n_k \approx$ a few hundreds for AAs, 1000-2000 bp for DNA sequences.

• DP on the full $k$-dim box of volume $n_1 \times n_2 \times \ldots \times n_k$ takes $O(n_1 \cdot n_2 \cdot \ldots \cdot n_k \cdot 2^k)$.

• Such running time is very slow even for $k = 3$, and totally infeasible for $k \geq 6$.

• Certain versions of MSA are known to be NPH, so an exact poly time (poly in what?) is unlikely.
Coping with NP hardness

• Several (of many) approaches:
Coping with NP hardness

• Several (of many) approaches:
  • Heuristics.
Coping with NP hardness

- Several (of many) approaches:
- Hueristics.
- Fixed parameter complexity.
Coping with NP hardness

• Several (of many) approaches:
  • Hueristics.
  • Fixed parameter complexity.
  • Poly time approximation algorithms.
MSA Hueristics

• Several (of many) approaches:
MSA Hueristics

• Several (of many) approaches:
• Searching only a promising subset of the $k$-dim box (used by Carillo-Lipman).

© Benny Chor
MSA Hueristics

• Several (of many) approaches:
• Searching only a **promising subset** of the $k$-dim box (used by Carillo-Lipman).
• Progressive pairwise alignment (used in **CLUSTALW**, a very popular package.)
MSA Hueristics

- Several (of many) approaches:
- Searching only a **promising subset** of the $k$-dim box (used by Carillo-Lipman).
- Progressive pairwise alignment (used in **CLUSTALW**, a very popular package.)
Fixed Parameter Complexity

- Is MSA in FTP?
Approximation Algorithms

- Gusfield 2-approximation algorithm.
Approximation Algorithms

- Gusfield 2-approximation algorithm.
- Setting: Sum of pairs. Distance Measure.
Dry Run of 2 Approx. MSA

• Four Sequences: $S_1, S_2, S_3, S_4$. 

$d(mismatch) = 3$, $d(indel) = 2$ (triangle inequality holds). 

$S_1 = \text{AAAAAAAAAA}$ 

$S_2 = \text{AAAAAAAAAA}$ 

$S_3 = \text{AAAAA}$ 

$S_4 = \text{TAAAAAAAAAA}$ 

$d(S_1, S_i) = 3$, $d(S_i, S_j) = 4$. 

So $S_1$ is the “center.”
Dry Run of 2 Approx. MSA

• Four Sequences: $S_1, S_2, S_3, S_4$.
• $d$(mismatch) = 3, $d$(indel) = 2
  (triangle inequality holds).
Dry Run of 2 Approx. MSA

• Four Sequences: $S_1, S_2, S_3, S_4$.
• $d$(mismatch) = 3, $d$(indel) = 2
  (triangle inequality holds).
Dry Run of 2 Approx. MSA

- Four Sequences: $S_1, S_2, S_3, S_4$.
- $d(\text{mismatch}) = 3$, $d(\text{indel}) = 2$ (triangle inequality holds).
- $S_1 = \text{AAAAAAAAAAAAA}$
  $S_2 = \text{AAAAAAAAAAAAA}\text{C}$
  $S_3 = \text{AAAAAAAAAGAAAAA}$
  $S_4 = \text{TAAAAAAAAA}$
Dry Run of 2 Approx. MSA

• Four Sequences: $S_1, S_2, S_3, S_4$.

• $d(\text{mismatch}) = 3$, $d(\text{indel}) = 2$
  (triangle inequality holds).

• $S_1 = \text{AAAAAAAAAAA}$
  $S_2 = \text{AAAAAAAAAAAC}$
  $S_3 = \text{AAAAAAAAAGAAAAA}$
  $S_4 = \text{TAAAAAAAAAAAA}$

• $d(S_1, S_i) = 3$, $d(S_i, S_j) = 4$. 
Dynamics of Progressive MSA
Dynamics of Progressive MSA

1) \( S_1 = \text{AAAAAA} - \text{AAAA} \)
\( S_2 = \text{AAAAAA} \text{AAAAA} \text{AAC} \)
Dynamics of Progressive MSA

1) \[ S_1 = \text{AAAAAAAAAA} - \]
   \[ S_2 = \text{AAAAAAAAAAAC} \]
Dynamics of Progressive MSA

1) \( S_1 = \text{AAAAAAAAAAAAA} - S_2 = \text{AAAAAAAAAAAAAC} \)

2) \( S_1 = \text{AAAAAAA} - \text{AAAAA} - S_2 = \text{AAAAAAA} - \text{AAAAAAC} \)
   \( S_3 = \text{AAAAAAAGAAAAA} - \)
Dynamics of Progressive MSA

1) \( S_1 = \text{AAAAAAAAAAAAAAAA} \)  
   \( S_2 = \text{AAAAAAAAAAAAAC} \)

2) \( S_1 = \text{AAAAAAA} \)  
   \( S_2 = \text{AAAAAAA} \)  
   \( S_3 = \text{AAAAAAAGAAAAAAAA} \)
Dynamics of Progressive MSA

1) \( S_1 = \text{AAAAAAAAA} - \)  
   \( S_2 = \text{AAAAAAAAAAC} \)

2) \( S_1 = \text{AAAAAAA} - \text{AAAAAA} - \)  
   \( S_2 = \text{AAAAAAA} - \text{AAAAAAC} \)  
   \( S_3 = \text{AAAAAAAAGAAAAA} - \)  

3) \( S_1 = -\text{AAAAAAA} - \text{AAAAAA} - \)  
   \( S_2 = -\text{AAAAAAA} - \text{AAAAAAC} \)  
   \( S_3 = -\text{AAAAAAAAGAAAAA} - \)  
   \( S_4 = \text{TAAAAAAA} - \text{AAAAAA} - \)
Dynamics of Progressive MSA

1) \( S_1 = \text{AAAAAAAAAAAAA} - \)
   \( S_2 = \text{AAAAAAAAAAAAA}C \)

2) \( S_1 = \text{AAAAAAA} - \text{AAAAA} - \)
   \( S_2 = \text{AAAAAAA} - \text{AAAAAAC} \)
   \( S_3 = \text{AAAAAAAGAAAAA} - \)

3) \( S_1 = \text{AAAAAAAA} - \text{AAAAAA} - \)
   \( S_2 = \text{AAAAAAAA} - \text{AAAAAAC} \)
   \( S_3 = \text{AAAAAAA}GAAAAA - \)
   \( S_4 = \text{TAAAAAAA} - \text{AAAAA} - \)
General Analysis of MSA Quality

- Denote $D(S_i, S_j) = \text{MSA--induced distance of } S_i, S_j$. 
General Analysis of MSA Quality

- Denote $D(S_i, S_j) = \text{MSA–induced distance of } S_i, S_j$.
- Denote $d(S_i, S_j) = \text{optimal (pairwise–induced) distance of } S_i, S_j$. 
General Analysis of MSA Quality

• Denote $D(S_i, S_j) =$ MSA–induced distance of $S_i, S_j$.

• Denote $d(S_i, S_j) =$ optimal (pairwise–induced) distance of $S_i, S_j$.

• By algorithm, $D(S_1, S_j) = d(S_1, S_j)$. 
More Analysis of MSA Quality

• By triangle inequality,

\[ D(S_i, S_j) \leq D(S_i, S_1) + D(S_1, S_j) \]

\[ = d(S_i, S_1) + d(S_1, S_j) \]
More Analysis of MSA Quality

• By triangle inequality,

\[ D(S_i, S_j) \leq D(S_i, S_1) + D(S_1, S_j) \]
\[ = d(S_i, S_1) + d(S_1, S_j) \]

• Thus

\[ \sum_{i=1}^{k} \sum_{j=i+1}^{k} D(S_i, S_j) \leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \]
More Analysis of MSA Quality

- By triangle inequality,

\[ D(S_i, S_j) \leq D(S_i, S_1) + D(S_1, S_j) \]
\[ = d(S_i, S_1) + d(S_1, S_j) \]

- Thus

\[ \sum_{i=1}^{k} \sum_{j=i+1}^{k} D(S_i, S_j) \leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \]

- Standard arithmetic now implies 2 approx. ratio (in fact \(2 - 1/k\)).
More Analysis of MSA Quality

• By triangle inequality,

\[ D(S_i, S_j) \leq D(S_i, S_1) + D(S_1, S_j) = d(S_i, S_1) + d(S_1, S_j) \]

• Thus

\[ \sum_{i=1}^{k} \sum_{j=i+1}^{k} D(S_i, S_j) \leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \]

• Standard arithmetic now implies 2 approx. ratio (in fact \(2 - 1/k\)).

• See Gusfield for full details.