Computational Genomics (0382.3102)

Lecture 5

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 $S_1, S_2, \ldots, S_k$, $k \geq 3$.

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:
- Hueristics.

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).

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.

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).

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).

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).
- $S_1 =$AAAAAAAAAAA
$S_2 =$AAAAAAAAAAAC
$S_3 =$AAAAAAGAAAAA
$S_4 =$TAAAAAAAAAAA

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).
- $S_1 =$AAAAAAAAAAA
  $S_2 =$AAAAAAAAAAAC
  $S_3 =$AAAAAAGAAAAA
  $S_4 =$TAAAAAAAAAAA
- $d(S_1, S_i) = 3$, $d(S_i, S_j) = 4$.

Dynamics of Progressive MSA

Dynamics of Progressive MSA

1) $S_1 =$ AAAAAAAAAAA-
$S_2 =$ AAAAAAAAAAAC

Dynamics of Progressive MSA

1) $S_1 =$ AAAAAAAAAAA-
$S_2 =$ AAAAAAAAAAAC

Dynamics of Progressive MSA

1) $S_1 =$ `AAAAAAAAAAA-`
$S_2 =$ `AAAAAAAAAAAC`

2) $S_1 =$ `AAAAAA-AAAAA-`
$S_2 =$ `AAAAAA-AAAAAC`
$S_3 =$ `AAAAAAGAAAAA-`

Dynamics of Progressive MSA

1) $S_1 =$ AAAAAAAAAAA-
$S_2 =$ AAAAAAAAAAAC

2) $S_1 =$ AAAAAA-AAAAA-
$S_2 =$ AAAAAA-AAAAAC
$S_3 =$ AAAAAAGAAAAA-

Dynamics of Progressive MSA

1) $S_1 =$ AAAAAAAAAAA-
$S_2 =$ AAAAAAAAAAAC

2) $S_1 =$ AAAAAA-AAAAA-
$S_2 =$ AAAAAA-AAAAAC
$S_3 =$ AAAAAAGAAAAA-

3) $S_1 =$ -AAAAAA-AAAAA-
$S_2 =$ -AAAAAA-AAAAAC
$S_3 =$ -AAAAAAGAAAAA-
$S_4 =$ TAAAAAA-AAAAA-

Dynamics of Progressive MSA

1) $S_1 =$ `AAAAAAAAAAA-`
$S_2 =$ `AAAAAAAAAAAC`

2) $S_1 =$ `AAAAAA-AAAAA-`
$S_2 =$ `AAAAAA-AAAAAC`
$S_3 =$ `AAAAAAGAAAAA-`

3) $S_1 =$ `-AAAAAA-AAAAA-`
$S_2 =$ `-AAAAAA-AAAAAC`
$S_3 =$ `-AAAAAAGAAAAA-`
$S_4 =$ `TAAAAAA-AAAAA-`

General Analysis of MSA Quality

- Denote $D(S_i, S_j) =$ MSA–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.

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,

$$\begin{aligned} D(S_i, S_j) &\leq D(S_i, S_1) + D(S_1, S_j) \\ &= d(S_i, S_1) + d(S_1, S_j) \end{aligned}$$

More Analysis of MSA Quality

- By triangle inequality,

$$\begin{aligned} D(S_i, S_j) &\leq D(S_i, S_1) + D(S_1, S_j) \\ &= d(S_i, S_1) + d(S_1, S_j) \end{aligned}$$

- Thus

$$\begin{aligned} \sum_{i=1}^{k} \sum_{j=i+1}^{k} & D(S_i, S_j) \\ \leq & \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) \\ & + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \end{aligned}$$

More Analysis of MSA Quality

- By triangle inequality,

$$\begin{aligned} D(S_i, S_j) &\leq D(S_i, S_1) + D(S_1, S_j) \\ &= d(S_i, S_1) + d(S_1, S_j) \end{aligned}$$

- Thus

$$\begin{aligned} \sum_{i=1}^{k} \sum_{j=i+1}^{k} & D(S_i, S_j) \\ &\leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) \\ &\quad + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \end{aligned}$$

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

More Analysis of MSA Quality

- By triangle inequality,

$$\begin{aligned} D(S_i, S_j) &\leq D(S_i, S_1) + D(S_1, S_j) \\ &= d(S_i, S_1) + d(S_1, S_j) \end{aligned}$$

- Thus

$$\begin{aligned} \sum_{i=1}^{k} \sum_{j=i+1}^{k} & D(S_i, S_j) \\ \leq & \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_i, S_1) \\ & + \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(S_1, S_j) \end{aligned}$$

- Standard arithmetic now implies 2 approx. ratio (in fact $2 - 1/k$).
- See Gusfield for full details.