Approximation of RNA Multiple Structural Alignment

Marcin Kubica[1], Romeo Rizzi[2], Stéphane Vialette[3] and Tomasz Waleń[1]

[1]Faculty of Mathematics, Informatics and Applied Mathematics
Warsaw University, Poland

[2]Dipartimento di Matematica ed Informatica (DIMI),
Università di Udine, Via delle Scienze 208, I-33100 Udine, Italy

[3]Laboratoire de Recherche en Informatique (LRI), UMR CNRS 8623
Faculté des Sciences d'Orsay - Université Paris-Sud, 91405 Orsay, France

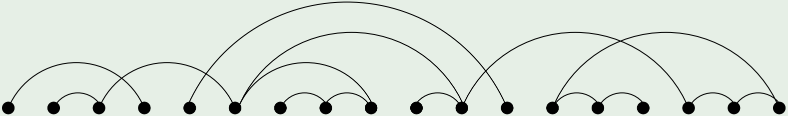
CPM, 2006-07-06

Linear graph

Definition

A linear graph of order n is a vertex-labeled graph where each vertex is labeled by a distinct label from $\{1, 2, \ldots, n\}$.

Example

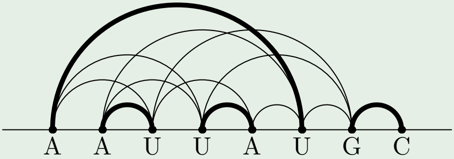


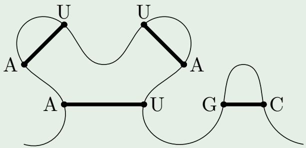
From ncRNA to linear graphs

Definition

- nucleotides are represented by vertices,
- possible bonds between nucleotides are represented by edges,
- non–crossing subset of edges represent possible folding

Example



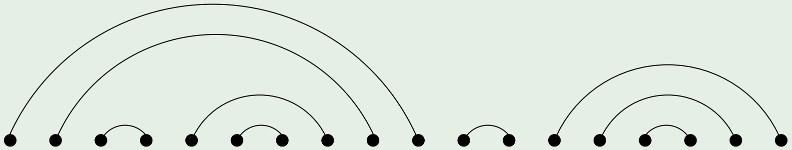


Linear graph

Definition

A linear graph is nested if no two edges cross.

Example



The Max-NLS problem

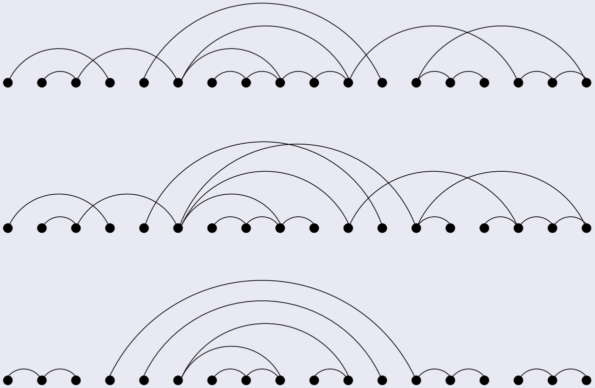
Let $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ be a set of linear graphs.
Find a maximum size common nested linear subgraph of $G_i \in \mathcal{G}$.

Example

The Max-NLS problem

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ be a set of linear graphs.
Find a maximum size common nested linear subgraph of $G_i \in \mathcal{G}$.

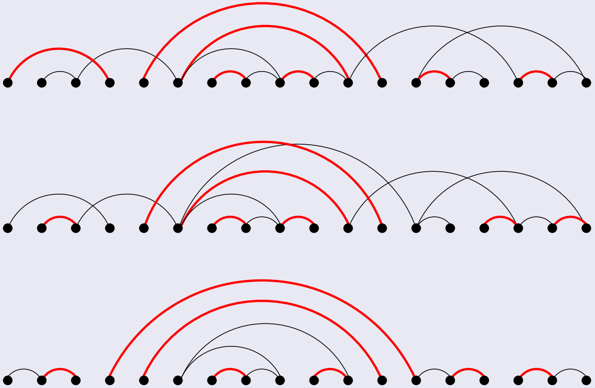
Example



The Max-NLS problem

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ be a set of linear graphs.
Find a maximum size common nested linear subgraph of $G_i \in \mathcal{G}$.

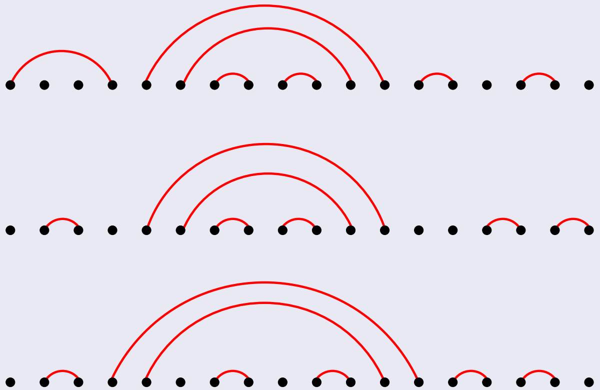
Example



The Max-NLS problem

Let $\mathcal{G} = \{G_1, G_2, \ldots, G_k\}$ be a set of linear graphs.
Find a maximum size common nested linear subgraph of $G_i \in \mathcal{G}$.

Example

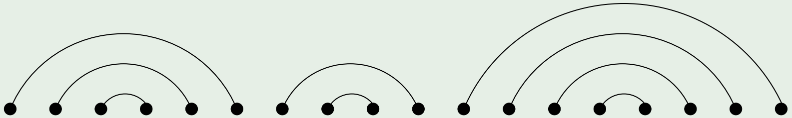


Flat linear graph

Definition

A nested linear graph is flat if it contains no branching edges, *i.e.*, it is composed of an ordered set of stacks.

Example

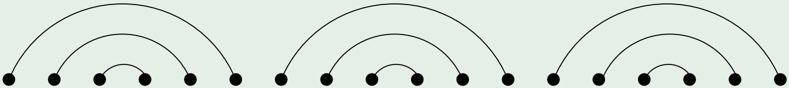


Level linear graph

Definition

A flat linear graph is level if it is composed of an ordered set of stacks of the same height.

Example



Approximation of MAX-NLS with MAX-LLS

Theorem (Davydov, Batzoglou, 2004)

The MAX-NLS problem is approximable within ratio $O(\log^2 m_{opt})$. Where m_{opt} is the maximum number of edges of an optimal solution.

Comments

$$\boxed{\text{MAX-NLS}} \xrightarrow[\times \log m_{opt}]{} \boxed{\text{MAX-FLS}} \xrightarrow[\times \log m_{opt}]{} \boxed{\text{MAX-LLS}}$$

Approximation of MAX-NLS with MAX-LLS

Theorem

The MAX-NLS problem is approximable within ratio $O(\log m_{opt})$.
Where m_{opt} is the maximum number of edges of an optimal solution.

Comments

$$\boxed{\text{MAX-NLS}} \quad \xrightarrow[\times \log m_{opt}]{} \quad \boxed{\text{MAX-LLS}}$$

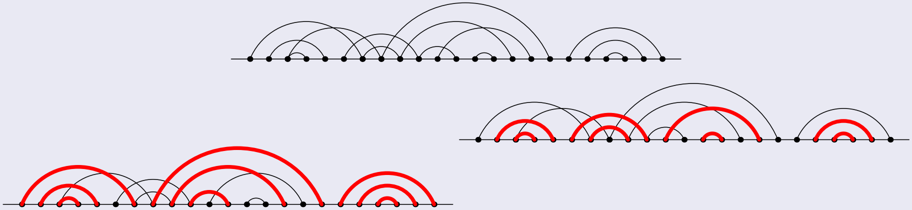
The $O(\log m)$ approximation bound is tight.

Level signature

Definition

Level signature of G is a function such, that:
(i) $s(h)$ is the maximum width of a level subgraph of G with height h;
(ii) if G has no level subgraph of height h, then $s(h) = 0$.

Example



Maximum level subgraphs of G with height 3 (on the left), and height 2 (on the right). The level signature of the graph is: $s(1) = 5$, $s(2) = 4$, $s(3) = 3$, $s(4) = 0$.

Approximation of MAX-NLS with MAX-LLS

Theorem (Davydov, Batzoglou, 2004)

The MAX-LLS problem is solvable in $O(k \cdot n^5)$ time.

Theorem

The MAX-LLS problem is solvable in $O(k \cdot n^2)$ time.

Outline

1. compute signatures of each graph (dynamic programming),
2. compute common signature,
3. choose best solution.

Approximation of MAX-NLS with MAX-LLS

Theorem (Davydov, Batzoglou, 2004)

The MAX-LLS problem is solvable in $O(k \cdot n^5)$ time.

Theorem

The MAX-LLS problem is solvable in $O(k \cdot n^2)$ time.

Outline

1. compute signatures of each graph (dynamic programming),
2. compute common signature,
3. choose best solution.

A polynomial-time algorithm for fixed |G|

Theorem

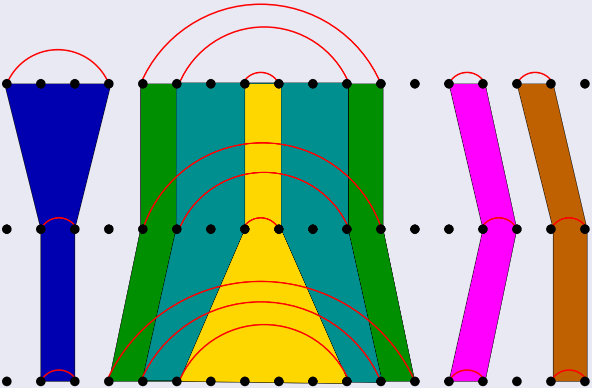
The Max-NLS *problem is solvable in* $\mathcal{O}(m^{2k} \cdot \log^{k-2} m^k \cdot \log\log m^k)$ *time, where* $k = |\mathcal{G}|$ *and* $m = \max\{|\mathbf{E}(G_i)| : G_i \in \mathcal{G}\}$.

Comments

- Geometric representation of linear graphs: d-trapezoids
- Max weighted Independent Set in d-trapezoid graphs.
- Dynamic programming

MAX-NLS and d–trapezoids

Example



Hardness results

Theorem (Davydov, Batzoglou. 2004)

The Max-NLS *problem is* **NP***-complete.*

Theorem

The Max-NLS *problem for flat linear graphs of height at most 2 is* **NP***-complete.*

Hardness results

Theorem (Davydov, Batzoglou. 2004)

The Max-NLS *problem is* **NP***-complete.*

Theorem

The Max-NLS *problem for flat linear graphs of height at most 2 is* **NP***-complete.*

MAX-NLS Problem for ncRNA Generated Linear Graphs

Restricted linear graphs

Graphs produced from the sequences using simple rules.

$$(i, j) \in E \text{ iff character } S[i] \text{ matches } S[j]$$

Results

- For any finite fixed alphabet we can approximate MAX-NLS with $O(1)$ approximation factor, in $O(n \cdot k)$ time
- For ncRNA we can show that the approximation factor is not greater than $\frac{1}{4}$.

MAX-NLS Problem for ncRNA Generated Linear Graphs

Restricted linear graphs

Graphs produced from the sequences using simple rules.

$$(i, j) \in E \text{ iff character } S[i] \text{ matches } S[j]$$

Results

- For any finite fixed alphabet we can approximate MAX-NLS with $O(1)$ approximation factor, in $O(n \cdot k)$ time
- For ncRNA we can show that the approximation factor is not greater than $\frac{1}{4}$.

Conclusions

- Faster MAX-NLS/MAX-LLS approximation algorithm $O(k \cdot n^2)$
- Better approximation ration proved $O(\log m_{opt})$
- Exact algorithm for MAX-NLS running in $\mathcal{O}(m^{2k} \cdot \log^{k-2} m^k \cdot \log\log m^k)$ time
- Improved hardness results
- $O(1)$ MAX-NLS approximation algorithm for a finite fixed alphabet of nucleotides, running in $O(n \cdot k)$ time
- $\frac{1}{4}$ MAX-NLS approximation algorithm for ncRNA derived linear graphs