Common Structured Patterns in Linear Graphs: Approximation and Combinatorics

Guillaume Fertin, Danny Hermelin,
Romeo Rizzi, and Stéphane Vialette

CPM, July 9 to 11, 2007

Outline

Context

Davydov and Batzoglou

- Framework for analysing the problem of ncRNA multiple structural alignment:

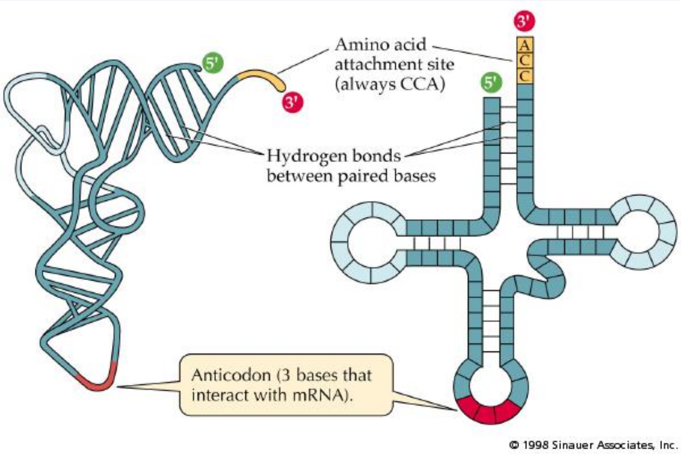
  "**finding the largest secondary structure of an ncRNA.**"

- Graph theoretic formulation:

  "**finding the largest non-crossing subgraph in the linear graph derived from the sequence.**"

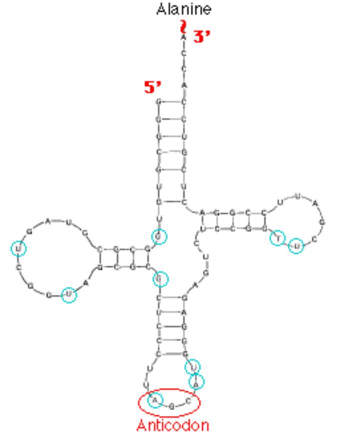
RNA secondary structure

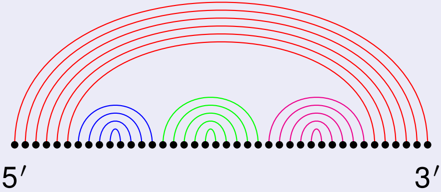
Transfer RNA



From RNA to linear graphs

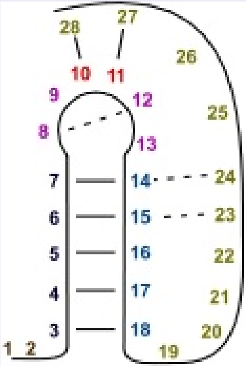
tRNA secondary structure

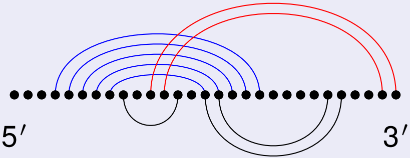




From RNA to linear graphs

A RNA pseudoknot (437D)



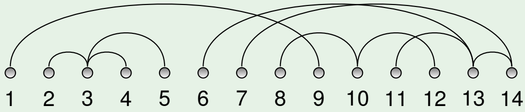


Linear graphs and patterns

Definition

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

Example

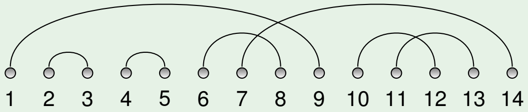


Linear graphs and patterns

Definition

- Two edges of a graph are called *independent* if they do not share a vertex.
- A linear graph G is called *edge-independent* if it is composed of independent edges, *i.e.*, G is a matching.

Example

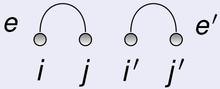


Relations between independent edges

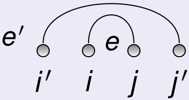
Definition

Let $e = (i, j)$ and $e' = (i', j')$ be two independent edges in a linear graph. We write

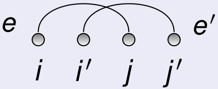
- $e < e'$ if $i < j < i' < j'$,



- $e \sqsubset e'$ if $i' < i < j < j'$,



- $e \between e'$ if $i < i' < j < j'$,



Comparability

Definition

Two (independent) edges e and e' are said to be R-comparable for some $R \in \{<, \sqsubset, \between\}$ if eRe' or $e'Re$.

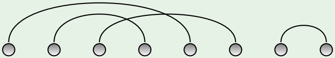
Definition

An edge-independent G is said to be *$\mathcal{M}$-comparable* for some non-empty $\mathcal{M} \subseteq \{<, \sqsubset, \between\}$ if any two distinct edges in G are R-comparable for some $R \in \mathcal{M}$.

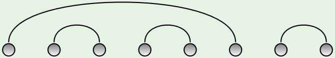
Some $\mathcal{M}$-comparable linear graphs

Example

$\mathcal{M} = \{<, \sqsubset, \between\}$

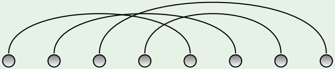


$\mathcal{M} = \{<, \sqsubset\}$

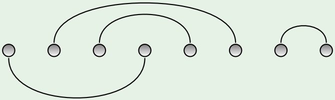


$\mathcal{M} = \{<, \between\}$

$\mathcal{M} = \{\sqsubset, \between\}$

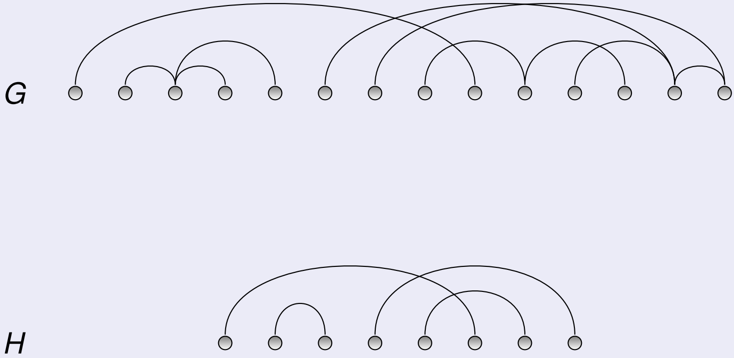


$\mathcal{M} = \{<, \sqsubset, \between\}$ and planar



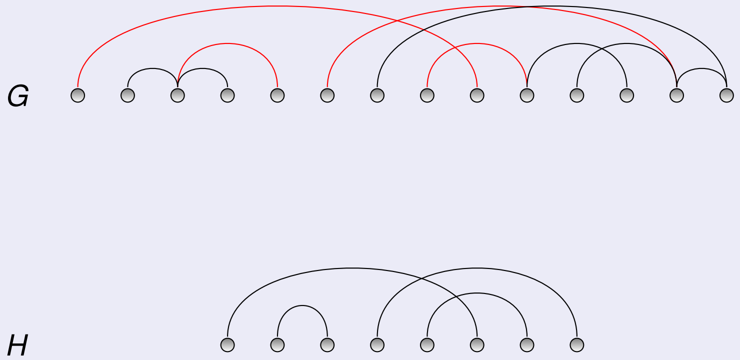
Occurrences in linear graphs

What is an occurrence of a pattern H in a linear graph G ?



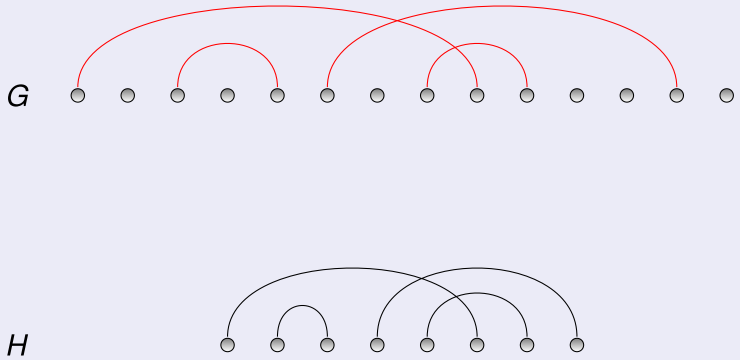
Occurrences in linear graphs

What is an occurrence of a pattern H in a linear graph G ?



Occurrences in linear graphs

What is an occurrence of a pattern *H* in a linear graph *G* ?



The problem we are interested in

$\mathcal{M}$-Common-Linear-Graph

- **Input** : A collection of n linear graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$, each of size at most m, and a model $\mathcal{M} \subseteq \{<, \sqsubset, \between\}$.
- **Solution** : A $\mathcal{M}$-comparable linear graph G_{sol} that occurs in each $G_i \in \mathcal{G}$.
- **Measure** : The size of the solution, *i.e.*, $|\mathbf{E}(G_{\text{sol}})|$.

Remarks

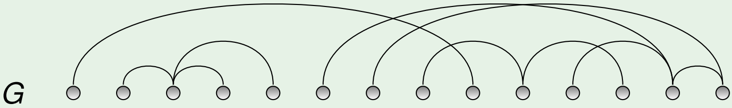
- Generalize Davydov and Batzoglou's framework [CPM'04].
- Intersection graph theory.
- Pattern matching for permutations.

Definitions

Parameters

- Width: Max. size of a $\{<\}$-comparable linear subgraph.
- Height: Max. size of a $\{\sqsubset\}$-comparable linear subgraph.
- Depth: Max. size of a $\{\between\}$-comparable linear subgraph.

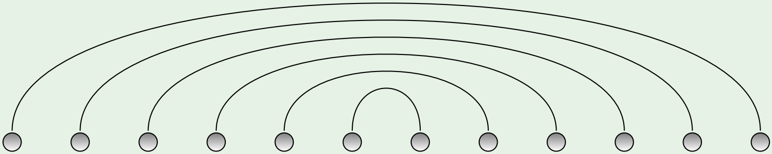
Example



$w(H) = 3, \quad h(G) = 2, \quad d(G) = 3$

Useful edge-disjoint linear graphs

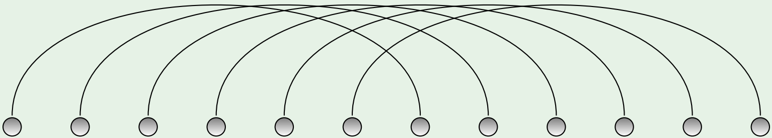
Example



A tower of height 6

Useful edge-disjoint linear graphs

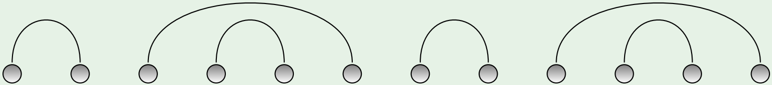
Example



A staircase of depth 6

Useful edge-disjoint linear graphs

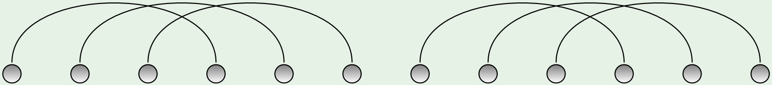
Example



A sequence of towers of width 4 and height 2

Useful edge-disjoint linear graphs

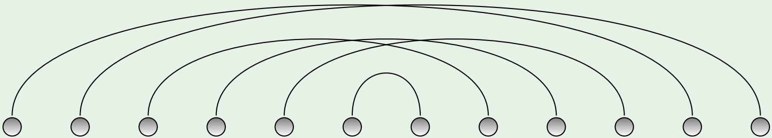
Example



A balanced sequence of staircases of width 2 and depth 3

Useful edge-disjoint linear graphs

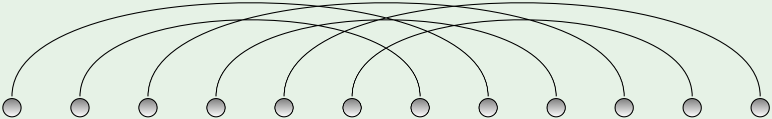
Example



A tower of staircases of height 3 and depth 3

Useful edge-disjoint linear graphs

Example



A balanced staircase of towers of height 2 and depth 3

Outline

Simple structured patterns

Theorem (Gupta, Lee and Leung, 1982)

The $\{<\}$-COMMON-LINEAR-GRAPH *problem is solvable in* $\mathcal{O}(m)$ *time, where* $m = |\mathbf{E}(G)|$.

Maximum cardinality pairwise disjoint subset of intervals.

Theorem (Chang and Wang, 1992)

The $\{\sqsubset\}$-COMMON-LINEAR-GRAPH *problem is solvable in* $\mathcal{O}(m \log\log m)$ *time, where* $m = |\mathbf{E}(G)|$.

Maximum cardinality pairwise nested subset of intervals.

Theorem (Tiskin, 2006)

The $\{\between\}$-COMMON-LINEAR-GRAPH *problem is solvable in* $\mathcal{O}(m^{1.5})$ *time, where* $m = |\mathbf{E}(G)|$.

Maximum cardinality pairwise overlapping subset of intervals.

Crossing-free patterns [CPM'04]

Theorem (Davydov and Batzoglou, 2004)

The $\{<, \sqsubset\}$-Common-Linear-Graph *problem is*

- **NP**–*complete even if each input linear graph* $G_i \in \mathcal{G}$ *is a matching, and*
- *approximable within ratio* $\mathcal{O}(\log^2 k)$, *where* k *is the size of an optimal solution.*

Keys ideas

- Balanced sequences of towers.
- $\mathcal{O}(nm^5)$ time algorithm, where $n = |\mathcal{G}|$ and m is the maximum size of a linear graph $G_i \in \mathcal{G}$.

Crossing-free patterns [CPM'06]

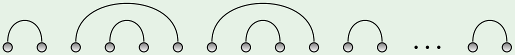
Theorem (Kubica, Rizzi, V and Waleń, 2006)

The $\{<, \sqsubset\}$-Common-Linear-Graph problem is

- **NP**–*complete even if each $G_i \in \mathcal{G}$ is a sequence of towers of height at most 2,*
- *solvable in $\mathcal{O}(m^{2n} \log^{n-2} m^n \log\log m^n)$ time, where m is the maximum size of an input graph $G_i \in \mathcal{G}$ and $n = |\mathcal{G}|$, and*
- *approximable within ratio $\mathcal{O}(\log k)$, where k is the size of an optimal solution.*

Example

G_i



Outline

$\mathcal{M} = \{<, \between\}$: Hardness result

Theorem (Li and Li, 2006)

The $\{<, \between\}$-COMMON-LINEAR-GRAPH *problem is* **NP***–complete even if* $|\mathcal{G}| = 2$.

Theorem

The $\{<, \between\}$-COMMON-LINEAR-GRAPH *problem is* **NP***–complete even if each* $G_i \in \mathcal{G}$ *is a sequence of staircases of depth at most* 2.

Remarks

- The problem is still open in case $|\mathcal{G}| = 1$.
- The problem is still open in case $|\mathcal{G}| = 2$ and each input linear graph is $\{<, \between\}$-comparable.

$\mathcal{M} = \{<, \between\}$: Approximation

Bal-Seq-Staircase

Balanced sequence of staircases of width w and depth d.

1. $E' \leftarrow \emptyset$.
2. **For** $i = 1$ to $m - 1$
   1. Let j be the smallest integer such that $G[i \dots j]$ contains as a subgraph a staircase of size d (set $j = \infty$ if no such integer exists).
   2. **if** $j \neq \infty$ **then** $E' \leftarrow E' \cup \{(i, j)\}$
3. Compute H, the maximum $\{<\}$-comparable subgraph of $G' = (V(G), E')$.
4. **if** $|E(H)| \geq w$ **then return** `true` **else return** `false`

Theorem

Algorithm Bal-Seq-Staircase(G, w, d) *runs in* $\mathcal{O}(m^{2.5} \log m)$ *time and returns* `true` *if and only if G contains a balanced sequence of staircases of width* w *and depth* d.

$\mathcal{M} = \{<, \between\}$: Approximation

Theorem

Any $\{<, \between\}$-matching of size k contains a balanced sequence of staircases of size at least $\frac{k}{2\,\mathcal{H}(k)}$.

Theorem

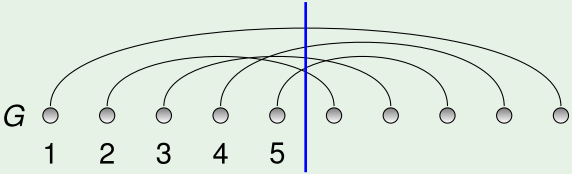
The $\{<, \sqsubset\}$-Common-Linear-Graph *problem is approximable within ratio* $2\mathcal{H}(k)$ *in* $\mathcal{O}(nm^{2.5}\log^2 m)$ *time, where k is the size of an optimal solution, $n = |\mathcal{G}|$, and m is the maximum size of a graph $G_i \in \mathcal{G}$.*

$\mathcal{M} = \{\sqsubset, \between\}$

Theorem

There is a bijection between $\{\sqsubset, \between\}$-comparable linear graphs of size n and permutations of size n.

Example



$\pi =$ 2 3 5 4 1

$\mathcal{M} = \{\sqsubset, \between\}$: Approximation

Theorem (Erdös, Szekeres, 1935)

The $\{\sqsubset, \between\}$-COMMON-MATCHING *problem is approximable within ratio* $\sqrt{k}$ *in* $\mathcal{O}(n\sqrt{m})$ *time, where* k *is the size of an optimal solution,* $n = |\mathcal{G}|$ *and* m *is the maximum size of an input graph* $G_i \in \mathcal{G}$*.*

Proof.

- Any permutation π of size k contains either an increasing or a decreasing subsequence of size $\sqrt{k}$.
- One can find in polynomial-time the largest $\{\sqsubset\}$-comparable or $\{\between\}$-comparable linear graph in a linear graph.

□

$\mathcal{M} = \{\sqsubset, \between\}$: Approximation

Theorem

There exists a $\{\sqsubset, \between\}$-comparable linear graph of size $\Omega(k^2)$ which does not contain a tower of staircases of size k.

- Towers of staircases cannot be used to obtain a much better approximation algorithm ...
- The above theorem can be modified to show that there exists a $\{\sqsubset, \between\}$-comparable linear graph of size $k = \Omega(k^2)$ which does not contain a tower of staircases, nor a staircase of towers, of size k.

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 1

Theorem

Let G be a $\{<, \sqsubset, \between\}$-comparable linear graph of size k. Then G contains a simple structured pattern of size at least $k^{1/3}$.

Theorem

Let k be an integer such that $k^{1/3}$ is also integer. Then there exists an $\{<, \sqsubset, \between\}$-comparable linear graph of size k that does not contain a simple structured pattern of size $\varepsilon\, k^{1/3}$ for any $\varepsilon > 1$.

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 1

Theorem

The $\{<, \sqsubset, \between\}$-COMMON-LINEAR-GRAPH problem is approximable within ratio $\mathcal{O}(k^{2/3})$ in $\mathcal{O}(nm^{1.5})$ time, where k is the size of an optimal solution, $n = |\mathcal{G}|$ and m is the maximum size of an input graph $G_i \in \mathcal{G}$.

Proof.

- Any edge-independent linear graph of size k contains a simple structured linear graph of size $k^{1/3}$.
- One can find the largest R-comparable pattern that occurs in a linear graph, $R \in \{<, \sqsubset, \between\}$, in $\mathcal{O}(nm^{1.5})$ time.

□

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 2

Theorem (Kostochka, Kratochvil, 1997)

Let G be a $\{<, \sqsubset, \between\}$-comparable linear graph of size k. Then G contains a subgraph of size $\Omega(\sqrt{k / \lg k})$ which is either $\{<, \sqsubset\}$-comparable or $\{\between\}$-comparable.

Note

A. Dumitrescu and G. Tóth, Ramsey-type Results for Unions of Comparability Graphs, *Graphs and Combinatorics*, 18 (2002), 245-251.

> *A graph of n vertices which is the union of two comparability graphs on the same vertex set, contains either a clique or an independent set of size at least $n^{1/3}$. Also, there exist such graphs for which the size of any clique or independent set is at most $n^{0.4118}$.*

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 2

Theorem

The $\{<, \sqsubset, \between\}$-Common-Linear-Graph *problem is approximable within ratio* $\mathcal{O}(\sqrt{k \log^3 k})$ *in* $\mathcal{O}(nm^2)$ *time, where* k *is the size of an optimal solution,* $n = |\mathcal{G}|$ *and* m *is the maximum size of an input graph* $G_i \in \mathcal{G}$*.*

Proof.

Two-step algorithm:

1. Maximum common $\{\between\}$-comparable linear subgraph.
2. Approximate maximum $\{<, \sqsubset\}$-comparable linear subgraph.

□

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 3

Theorem

Let G be a $\{<, \sqsubset, \between\}$-comparable linear graph of size k. Then G contains either a tower or a balanced sequence of staircases of size $\Omega(\sqrt{k/\lg k})$.

Proof.

- Consider the size of the maximum $\{\sqsubset\}$-comparable (resp. $\{<, \between\}$-comparable) linear subgraph of G.
- Apply Dilworth's theorem.
- Any $\{<, \between\}$-comparable linear graph of size k contains a balanced sequence of staircases of size at least $\dfrac{k}{2\,\mathcal{H}(k)}$.

□

$\mathcal{M} = \{<, \sqsubset, \between\}$: Approximation 3

Theorem

The $\{<, \sqsubset, \between\}$-COMMON-LINEAR-GRAPH problem is approximable within ratio $\mathcal{O}(\sqrt{k \log k})$ in $\mathcal{O}(nm^{2.5} \log^2 m)$ time, where k is the size of an optimal solution, $n = |\mathcal{G}|$ and m is the maximum size of an input graph $G_i \in \mathcal{G}$.

Proof.

- Find the maximum common tower in $\mathcal{G}$.
- Find the maximum balanced sequence of staircases in $\mathcal{G}$.

□

$\mathcal{M} = \{<, \sqsubset, \between\}$

Theorem

Let G be a $\{<, \sqsubset, \between\}$-comparable graph of size k. Then G has a subgraph of size $\varepsilon\, k^{2/3}$, where $\varepsilon = \frac{\sqrt{17}-1}{8}$, which is either $\{<, \sqsubset\}$-comparable, $\{<, \between\}$-comparable, or $\{\sqsubset, \between\}$-comparable.

Remarks

- The bound is probably not the best possible,
- but relatively tight: *Let k be an integer such that $\sqrt[3]{k}$ is integer. Then there exists a $\{<, \sqsubset, \between\}$-comparable linear graph of size k that contains neither a $\{<, \sqsubset\}$-comparable subgraph, nor a $\{<, \between\}$-comparable subgraph, nor a $\{\sqsubset, \between\}$-comparable subgraph of size least $\varepsilon\, k^{2/3}$ for any $\varepsilon > 1$.*

Outline

1. Introduction
2. Known results
3. Our results
   - $\mathcal{M} = \{<, \between\}$
   - $\mathcal{M} = \{\sqsubset, \between\}$
   - $\mathcal{M} = \{<, \sqsubset, \between\}$
4. Open problems

Open problem [1]

$\mathcal{M} = \{<, \sqsubset\}$ and $\mathcal{M} = \{<, \between\}$

Improve the $\mathcal{O}(\log k)$ approximation ratio, where k is the size of an optimal solution:

- The $\{<, \sqsubset\}$-Common-Linear-Graph problem.
- The $\{<, \between\}$-Common-Linear-Graph problem.

Natural questions

- Approximable within a constant ratio ($\in$ **APX**) ?
- Impact of the structure of the input linear graphs $G_i \in \mathcal{G}$?

Open problem [2]

MODEL $= \{<, \between\}$

- A different strategy for finding a better aproximation ratio for the $\{<, \between\}$-COMMON-LINEAR-GRAPH problem for $|\mathcal{G}| = 2$ (or $|\mathcal{G}| = \mathcal{O}(1)$).
- Is the $\{<, \between\}$–SEARCH-LINEAR-GRAPH problem fixed-parameter tractable ?

Open problem [3]

$MODEL = \{\sqsubset, \between\}$

Find a better aproximation algorithm for the $\{\sqsubset, \between\}$-COMMON-LINEAR-GRAPH problem.

Remarks

- The (current) bottleneck for finding a better approximation for the $\{<, \sqsubset, \between\}$-COMMON-LINEAR-GRAPH problem.
- Contain a permutation problem as a special case.
- We believe the ratio $k^{1/2}$ to be the best possible.

Open problem [4]

$MODEL = \{\sqsubset, \between\}$

Given a two $\{\sqsubset, \between\}$-comparable linear graphs G and H of size m and k, respectively, is there an algorithm to find an occurrence of H in G that runs in $f(k)\, m^c$ time, where f is an arbitrary function and c is a constant ?

Remarks

- Pattern matching for permutations and a special case of the $\{\sqsubset, \between\}$–SEARCH-LINEAR-GRAPH problem.
- Fixed-parameter tractability.
- The problem is **W[1]**–hard in case both G and H are $\{<, \sqsubset, \between\}$-comparable linear graphs.

APX–hardness for sequences of towers ?

Disjoint runs

4	1	3	2	6	0
2	1	4	2	7	9
4	3	1	6	1	2
3	9	2	5	7	1
3	3	3	2	1	5

2 2 5 $= 9$