Constrained RNA Folding - New Results

Marc Hellmuth   Daniel Merkle   Martin Middendorf

Faculty of Mathematics and Computer Science
University of Leipzig

February 20, 2007

Outline

Motivation

- RNA switch
  one RNA sequence $\Rightarrow$ different structures $\Rightarrow$ different functions
- Aptamer
  Aim: Finding a RNA sequence, which can fold into a special structure

QUESTION:

On what conditions is it possible to find a single RNA sequence which is compatible to given structures?

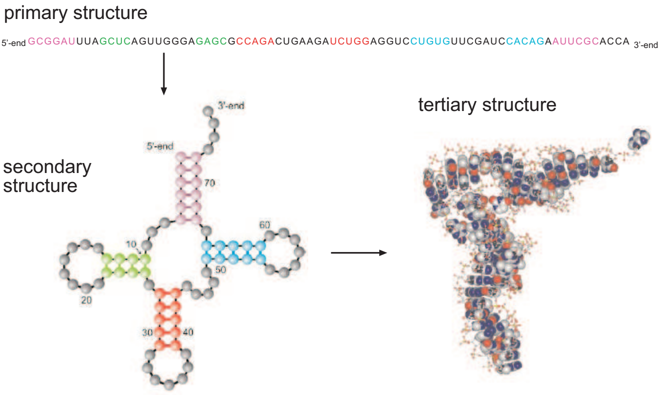
Motivation

- RNA switch
  one RNA sequence ⇒ different structures ⇒ different functions
- Aptamer
  Aim: Finding a RNA sequence, which can fold into a special structure

QUESTION:

On what conditions is it possible to find a single RNA sequence which is compatible to given structures?

RNA Structure



RNA Structure

Definition

A length n **RNA sequence** or **primary structure** is considered as a word s in the space $C_n = \{A, C, G, U\}^n$, i.e.

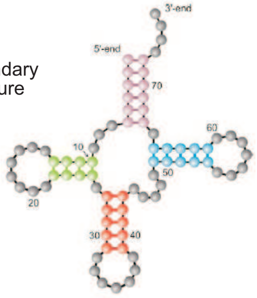
$$s = s_1 \dots s_n \in C_n = \{A, C, G, U\}^n.$$

RNA Structure

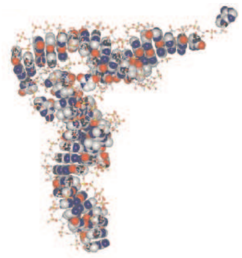
primary structure

5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end

secondary structure

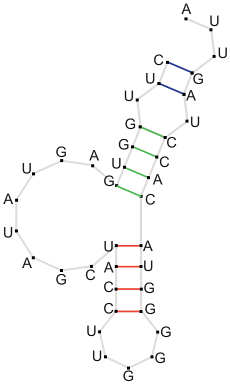


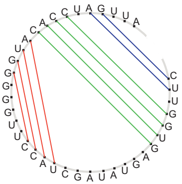
tertiary structure

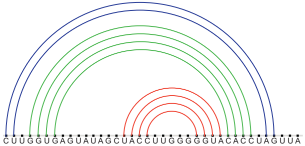


RNA Structure

A **shape** S of size n is a graph $S = (V_n, E)$ with set $V_n = \{v_1, \ldots, v_n\}$ of vertices and set E of independent edges such that for any two edges $\{v_i, v_j\}, \{v_k, v_l\} \in E$, where $i < j$ and $k < l$, it is not the case that $i < k < j < l$.

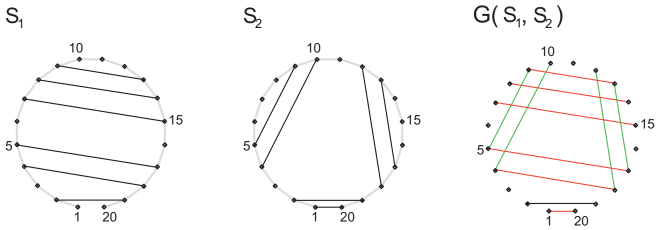






RNA Structure

For given shapes $S_1, \ldots, S_k$ of the same size n, we define the **graph of shapes** $G(S_1, \ldots, S_k)$ as $G(V_n, \cup_{i=1}^{k} E(S_i))$



Realization

On what conditions is it possible to find a single string, which is realizing any given shapes $S_1, \ldots, S_k$ of the same size?

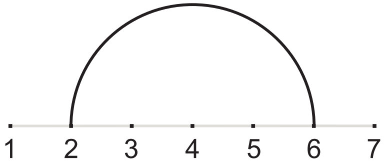
Definition

A string $s = s_1 \ldots s_n \in \{A, C, G, U\}^n$ **realizes** a shape S of size $n \Leftrightarrow \forall \{v_i, v_j\} \in E(S)$ holds:

$$s_i s_j \in \mathscr{B} = \{AU, UA, CG, GC, GU, UG\}.$$

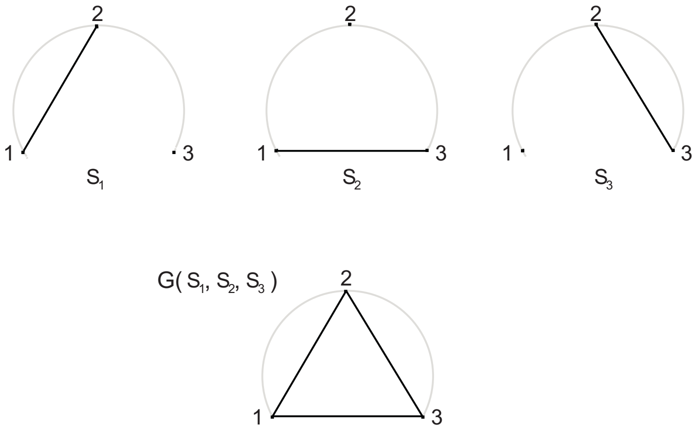
- $\mathscr{B}$ denotes the pairing rules.

Realization

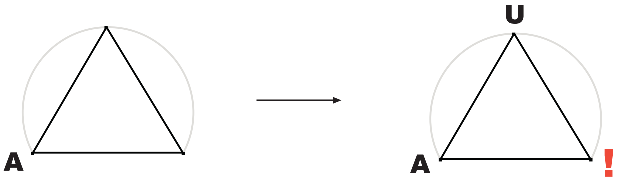


$s_2 s_6 \in \{AU, UA, CG, GC, GU, UG\}$

Realization



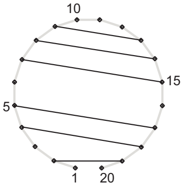
Realization



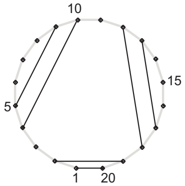
$$\mathscr{B} = \{AU, UA, CG, GC, GU, UG\}$$

Realization

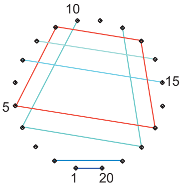
S_1



S_2



$G(S_1, S_2)$



Question:

On what conditions is it possible to find a single string, which is realizing any given shapes $S_1, \ldots, S_k$ of the same size?
If there is a single string, which is realizing all edges in the shapes $S_1, \ldots, S_k$, then this string realizes all the edges in the graph $G(S_1, \ldots, S_k)$.
Accordingly we can ask: What properties have to be fulfilled in the graph $G(S_1, \ldots, S_k)$, that there exists a single string, which is realizing all these shapes $S_1, \ldots, S_k$?

Question:

On what conditions is it possible to find a single string, which is realizing any given shapes $S_1, \ldots, S_k$ of the same size?
If there is a single string, which is realizing all edges in the shapes $S_1, \ldots, S_k$, then this string realizes all the edges in the graph $G(S_1, \ldots, S_k)$.
Accordingly we can ask: What properties have to be fulfilled in the graph $G(S_1, \ldots, S_k)$, that there exists a single string, which is realizing all these shapes $S_1, \ldots, S_k$?

Intersection Theorem

- $R(S)$ denotes the set of all sequences that are compatible with shape S.

1. **Intersection Theorem** [Reidys et al. 1995]
   For any two secondary structures S_1 and S_2 of same size holds: $R(S_1) \cap R(S_2) \neq \emptyset$
2. **Generalized Intersection Theorem** [Flamm et al. 2001]
   $\bigcap_{i=1}^{k} R(S_i) \neq \emptyset \Leftrightarrow$
   - $G(S_1, \ldots, S_k)$ is bipartite
   - there are no odd cycles in $G(S_1, \ldots, S_k)$
   - the shapes $S_1, \ldots, S_k$ are realizable by a single binary string, i.e. $\exists s \in \{0,1\}^n \Leftrightarrow \forall \{v_i, v_j\} \in E(S_l)$ holds: $s_i \neq s_j$

Intersection Theorem

- $R(S)$ denotes the set of all sequences that are compatible with shape S.

1. **Intersection Theorem** [Reidys et al. 1995]
   For any two secondary structures S_1 and S_2 of same size holds: $R(S_1) \cap R(S_2) \neq \emptyset$
2. **Generalized Intersection Theorem** [Flamm et al. 2001]
   $\bigcap_{i=1}^{k} R(S_i) \neq \emptyset \Leftrightarrow$
   - $G(S_1, \ldots, S_k)$ is bipartite
   - there are no odd cycles in $G(S_1, \ldots, S_k)$
   - the shapes $S_1, \ldots, S_k$ are realizable by a single binary string, i.e. $\exists s \in \{0,1\}^n \Leftrightarrow \forall \{v_i, v_j\} \in E(S_l)$ holds: $s_i \neq s_j$

Intersection Theorem

- $R(S)$ denotes the set of all sequences that are compatible with shape S.

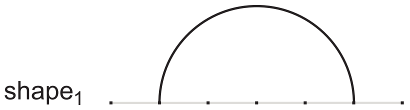
1. **Intersection Theorem** [Reidys et al. 1995]
   For any two secondary structures S_1 and S_2 of same size holds: $R(S_1) \cap R(S_2) \neq \emptyset$
2. **Generalized Intersection Theorem** [Flamm et al. 2001]
   $\bigcap_{i=1}^{k} R(S_i) \neq \emptyset \Leftrightarrow$
   - $G(S_1, \ldots, S_k)$ is bipartite
   - there are no odd cycles in $G(S_1, \ldots, S_k)$
   - the shapes $S_1, \ldots, S_k$ are realizable by a single binary string, i.e. $\exists s \in \{0,1\}^n \Leftrightarrow \forall \{v_i, v_j\} \in E(S_l)$ holds: $s_i \neq s_j$

Intersection Theorem

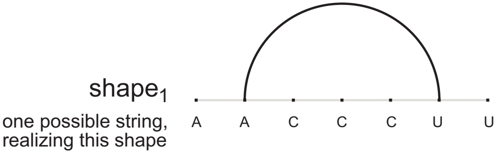
- $R(S)$ denotes the set of all sequences that are compatible with shape S.

1. **Intersection Theorem** [Reidys et al. 1995]
   For any two secondary structures S_1 and S_2 of same size holds: $R(S_1) \cap R(S_2) \neq \emptyset$
2. **Generalized Intersection Theorem** [Flamm et al. 2001]
   $\bigcap_{i=1}^{k} R(S_i) \neq \emptyset \; \Leftrightarrow$
   - $G(S_1, \ldots, S_k)$ is bipartite
   - there are no odd cycles in $G(S_1, \ldots, S_k)$
   - the shapes $S_1, \ldots, S_k$ are realizable by a single binary string, i.e. $\exists s \in \{0,1\}^n \Leftrightarrow \forall \{v_i, v_j\} \in E(S_l)$ holds: $s_i \neq s_j$

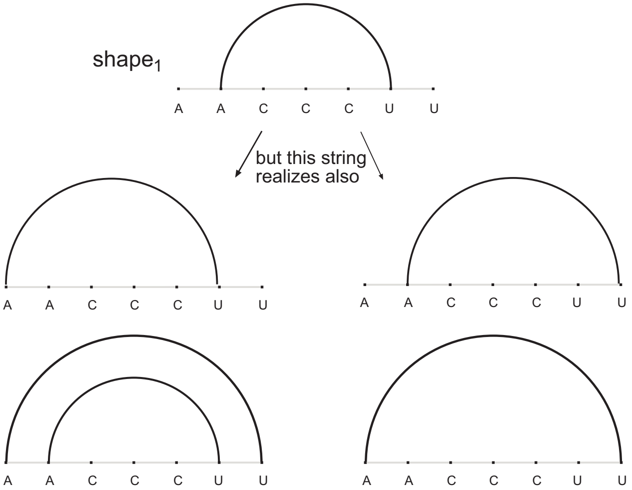
Why Extended Shapes?



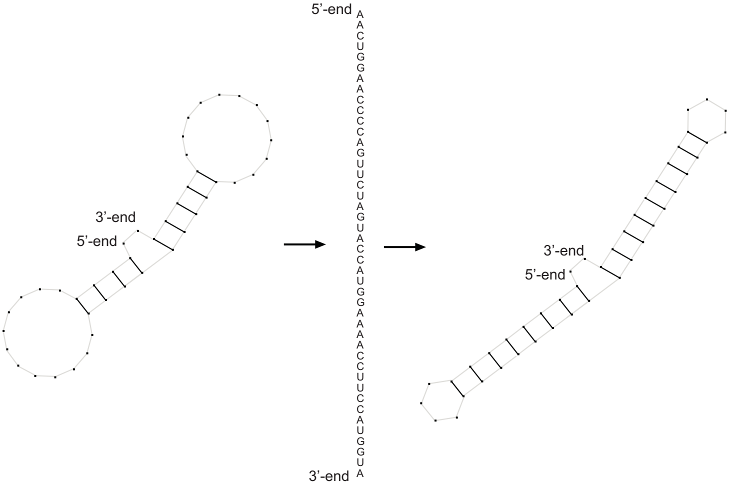
Why Extended Shapes?



Why Extended Shapes?



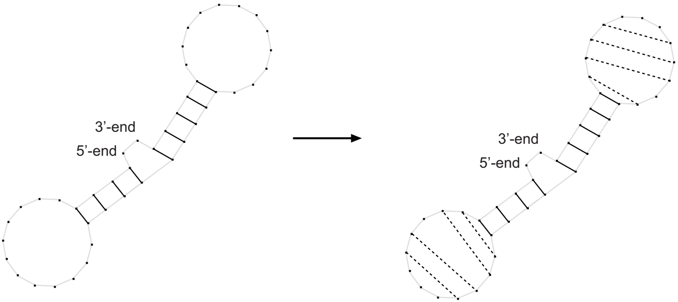
Why Extended Shapes?



Why Extended Shapes?

How can we find sequences, which are realizing given shapes, but also conserving the structure even better than the current sequences, in such a way that unpaired positions are considered?

Pseudo Edges



In this case:

- insert pseudo edges
- ask for additional conditions to a string, which is realising a shape with pseudo edges

Extended shapes

Definition

Let $S = G(V_n, E)$ be a shape of size n. The set $V'(S) = \{v_i \in V_n \mid \nexists v_j \in V_n,\ s.t.\{v_i, v_j\} \in E\}$ is the set of all vertices that are not paired.

- Now arbitrary vertices $v_i, v_j \in V'(S)$ will get connected by new edges $\{v_i, v_j\}$.
- These new edges are called **pseudo edges**.
- But connecting new vertices with pseudo edges is restricted to: **The characteristics of a shape have to be conserved.**

For a better understanding we are calling the 'old' edges: **regular edges**.

Extended shapes

Definition

Let $S = G(V_n, E)$ be a shape of size n. The set $V'(S) = \{v_i \in V_n \mid \nexists v_j \in V_n,\ s.t. \{v_i, v_j\} \in E\}$ is the set of all vertices that are not paired.

- Now arbitrary vertices $v_i, v_j \in V'(S)$ will get connected by new edges $\{v_i, v_j\}$.
- These new edges are called **pseudo edges**.
- But connecting new vertices with pseudo edges is restricted to: **The characteristics of a shape have to be conserved.**

For a better understanding we are calling the 'old' edges: **regular edges**.

Extended shapes

Definition

Let $\hat{E}(S)$ be the set of all inserted *pseudo edges* in S. We call the graph $\hat{S} = (V_n, E(S), \hat{E}(S)) := (V_n, \tilde{E} := E(S) \cup \hat{E}(S), \gamma)$, that emerges from the shape S, **shape extension** of S or also **extended shape**.

Extended shapes

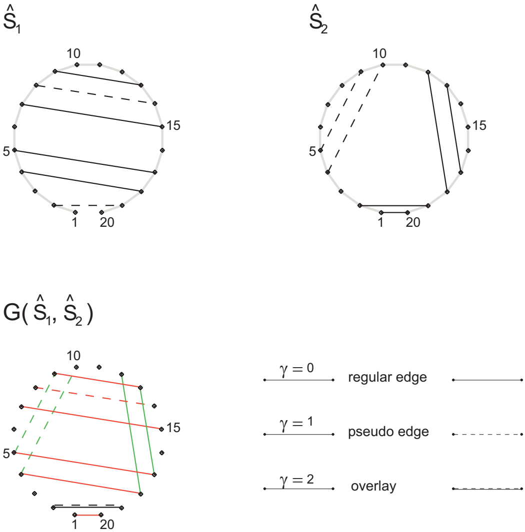
Definition

Let $\hat{S}_1 = (V_n, E(S_1), \hat{E}(S_1))$, ..., $\hat{S}_k = (V_n, E(S_k), \hat{E}(S_k))$ be extended shapes.

The **graph of extended shapes** $G(\hat{S}_1, \ldots, \hat{S}_k)$ is defined by:

$$(V, E, \hat{E}) := (V_n, \tilde{E} := \cup_{i=1}^{k} (E(S_i) \cup \hat{E}(S_i)), \gamma)$$

Extended shapes



Realizing Extended Shapes

For the sake of convenience we define the term *realizing an extended shape* at first for a binary alphabet.

Definition

A binary string $s = s_1 \ldots s_n \in \{0,1\}^n$ realizes an *extended shape* $\hat{S}$ of size $n \iff$

1. for all regular edges $\{v_i, v_j\} \in E(S)$ holds: $s_i \neq s_j$, i.e. $s_i s_j \in \mathscr{B} := \{01, 10\}$.
2. for all pseudo edges $\{v_i, v_j\} \in \hat{E}(S)$ holds: $s_i = s_j$, i.e. $s_i s_j \in \mathscr{B}_{PSE} := \{00, 11\}$.

Realizing Extended Shapes

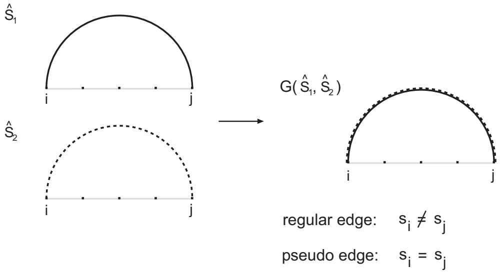
For the sake of convenience we define the term *realizing an extended shape* at first for a binary alphabet.

Definition

A binary string $s = s_1 \dots s_n \in \{0,1\}^n$ realizes an *extended shape* $\hat{S}$ of size $n \Longleftrightarrow$

1. for all regular edges $\{v_i, v_j\} \in E(S)$ holds: $s_i \neq s_j$, i.e. $s_i s_j \in \mathscr{B} := \{01, 10\}$.
2. for all pseudo edges $\{v_i, v_j\} \in \hat{E}(S)$ holds: $s_i = s_j$, i.e. $s_i s_j \in \mathscr{B}_{PsE} := \{00, 11\}$.

Realizing Extended Shapes



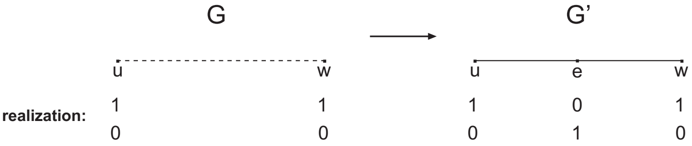
- Intersection Theorem isn't valid for extended shapes.

The same question, another answer

What properties have to be fulfilled in the graph of extended shapes $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. there exists a single binary string, which is realizing all these extended shapes $\hat{S}_1, \ldots, \hat{S}_k$

New Results

Let $G(\hat{S}_1, \hat{S}_2)$ be the graph of two extended shapes. Furthermore let G' be the graph, that emerges from $G(\hat{S}_1, \hat{S}_2)$ by replacing all pseudo edges $e = \{u, w\}$ with two new regular edges $\{u, e\}, \{e, w\}$ with a new vertex e not contained in V_n, i.e. deg(e)=2.



New Results

Let $G(\hat{S}_1, \hat{S}_2)$ be the graph of two extended shapes. Furthermore let G' be the graph, that arises from $G(\hat{S}_1, \hat{S}_2)$ by replacing all pseudo edges $e = \{u, w\}$ with two new regular edges $\{u, e\}, \{e, w\}$ with a new vertice e not contained in V_n, i.e. deg(e)=2.

Lemma

Two arbitrary extended shapes $\hat{S}_1, \hat{S}_2$ of size n can be realized by a single binary string $\Longleftrightarrow$ The graph G' is bipartite.

New Results

Theorem

Any extended shapes $\hat{S}_1, \ldots, \hat{S}_k$ of size n can be realized by a single binary string $\Leftrightarrow$

1. $E(S_l) \cap \hat{E}(S_m) = \emptyset$ *mit* $l \neq m \in \{1, \ldots, k\}$, *i.e. there are no overlays and*
2. *the graph* $G(\hat{S}_1, \ldots, \hat{S}_k)$ *does not contain cycles with an odd number of regular edges.*

New Results

Similar conditions to the RNA alphabet:

- for all regular edges $\{v_i, v_j\} \in E(S)$ holds: $s_i s_j \in \{AU, UA, CG, GC, GU, UG\}$
- for all pseudo edges $\{v_i, v_j\} \in \hat{E}(S)$ holds: $s_i = s_j$, i.e. $s_i s_j \in \{AA, UU, GG, CC\}$

New Results

Theorem

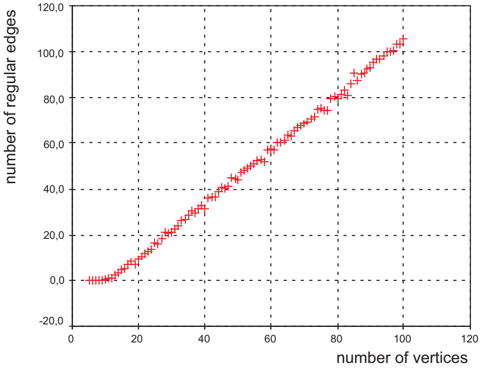
Any extended shapes $\hat{S}_1, \ldots, \hat{S}_k$ of size n can be realized by a single string $s \in \{A, C, G, U\}^n$ with the conditions above $\Leftrightarrow$ These extended shapes $\hat{S}_1, \ldots, \hat{S}_k$ of size n can be realized by a single binary string

New Results

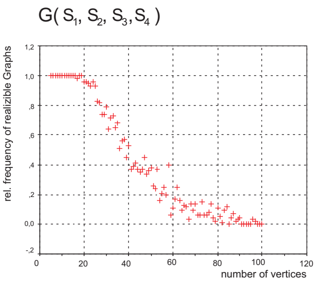
- If we have found sequences, that realize extended shapes,
- then we have found sequences, that realize the corresponding shapes,
- BUT also conserve the structure even better, in the case that unpaired positions are considered.

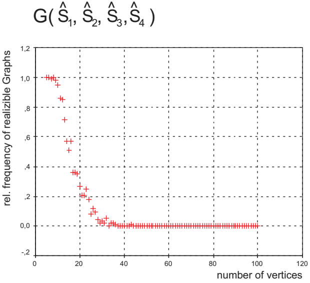
Empirical Results

$G(S_1, S_2, S_3, S_4)$

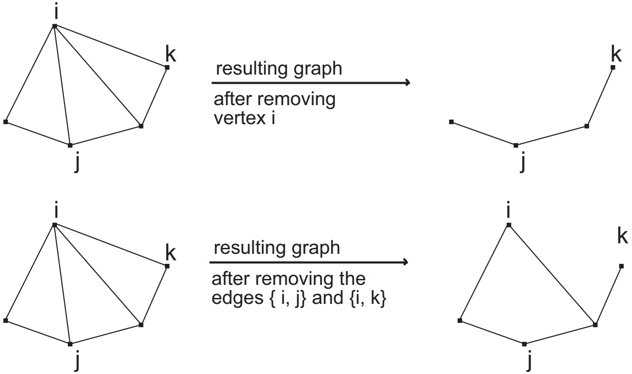


Empirical Results





Problems


i
k
j
resulting graph
after removing vertex i
k
j
i
k
j
resulting graph
after removing the edges { i, j} and {i, k}
i
k
j

Let $S_1, \ldots, S_k$ and $\hat{S}_1, \ldots, \hat{S}_k$ be any shapes or extended shapes respectively of the same size n.

Problem

Compute the minimum number of

1. VERTICES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
2. EDGES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
3. VERTICES, *that must be removed from the graph $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*
4. REGULAR AND PSEUDO EDGES, *that have to be removed from $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*

Let $S_1, \ldots, S_k$ and $\hat{S}_1, \ldots, \hat{S}_k$ be any shapes or extended shapes respectively of the same size n.

Problem

Compute the minimum number of

1. VERTICES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
2. EDGES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
3. VERTICES, *that must be removed from the graph $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*
4. REGULAR AND PSEUDO EDGES, *that have to be removed from $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*

Let $S_1, \ldots, S_k$ and $\hat{S}_1, \ldots, \hat{S}_k$ be any shapes or extended shapes respectively of the same size n.

Problem

Compute the minimum number of

1. VERTICES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
2. EDGES, *that must be removed from the graph $G(S_1, \ldots, S_k)$, s.t. $S_1, \ldots, S_k$ are realizable by a single string.*
3. VERTICES, *that must be removed from the graph $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*
4. REGULAR AND PSEUDO EDGES, *that have to be removed from $G(\hat{S}_1, \ldots, \hat{S}_k)$, s.t. $\hat{S}_1, \ldots, \hat{S}_k$ are realizable by a single string.*

Conclusion

1. we started with "normal" shapes
2. aim: finding sequences, that realizes shapes
3. we extended the shapes by inserting **pseudo edges**
4. aim: finding sequences, that realizes **extended shapes**.
5. There are NP-complete problems, arising from the term *realizing shapes and extended shapes*.

Thank you for your attention.