On the syntenic distance between genomes
(A few results and conjectures)

Laboratoire d'Algorithmique Combinatoire, Institut Pasteur
Inria, Lyon (from Sept. 1st)

This work with Nadia Pisanti (University of Pisa)

Genomes and genes

Genome

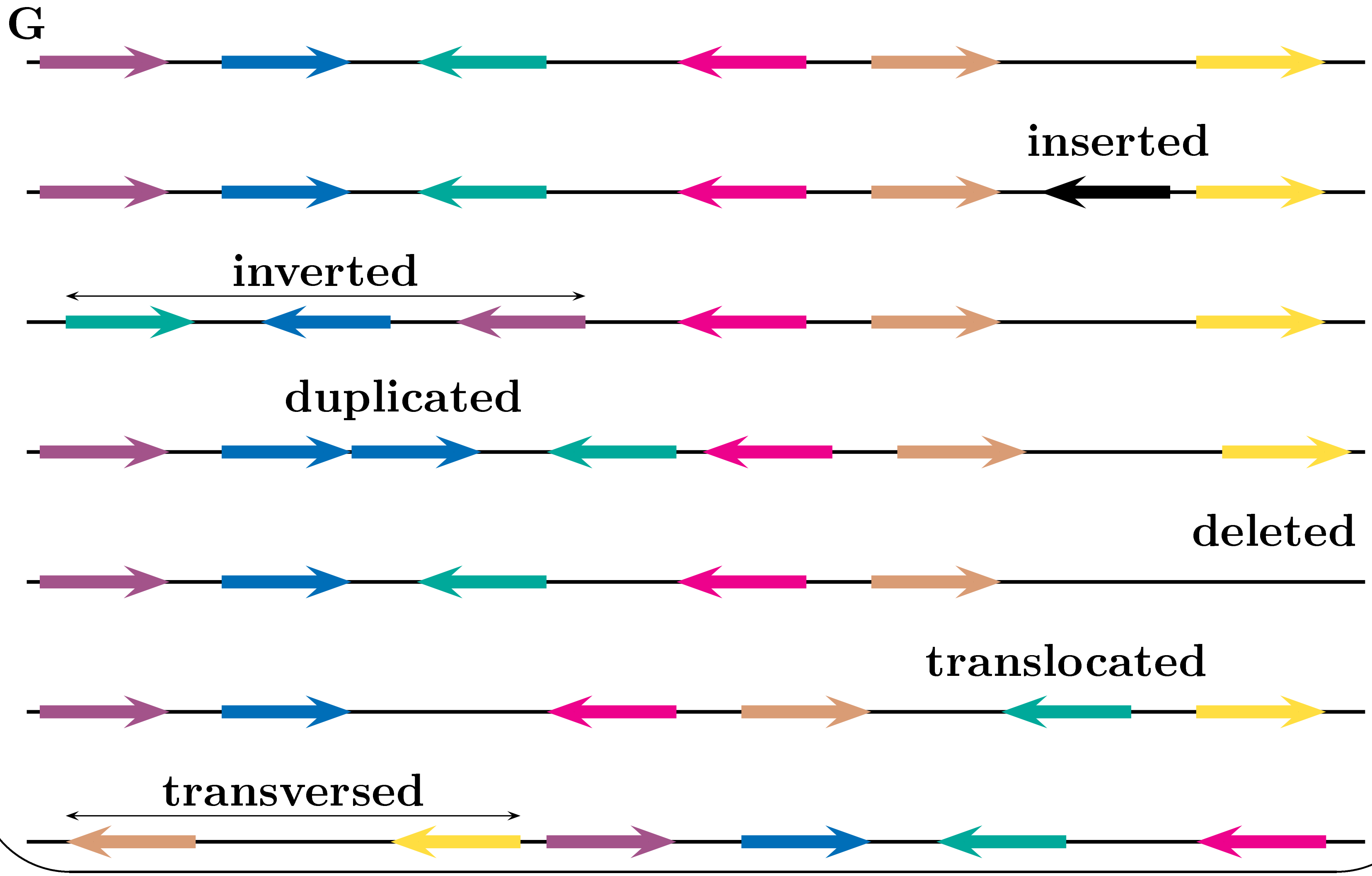
- **Full complement of genetic information in a cell**
- **Required for an organism to function**

Gene

- **Specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein**
- **Distributed among one (*e.g.* bacteria) or more (*e.g.* plants, mammals) chromosomes**

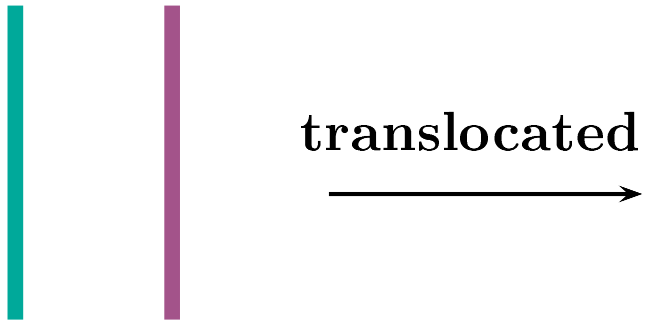
Genomes are dynamic objects

Segments (*e.g.* genes) are frequently



And, in the case of multichromosomal genomes,

Segments (*e.g.* genes, often various) are frequently



and chromosomes are frequently

Genome rearrangement distances

General rearrangement distance

The order of the genes along the genome is considered and any type of operation may be taken into account

Breakpoint distance

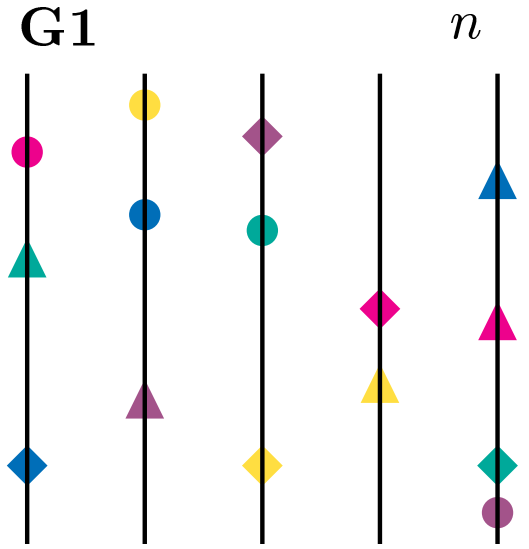
The order of the genes along the genome is considered but the type of operation is ignored

Syntenic distance

The order of the genes along the genome is ignored

This is meaningful only for multichromosomal genomes

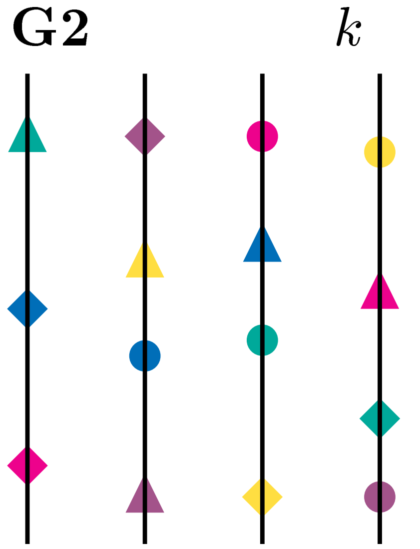
Syntenic distance



genome

=

set of
bag of genes



distance = minimum number of 3 types of operations to transform one genome into another

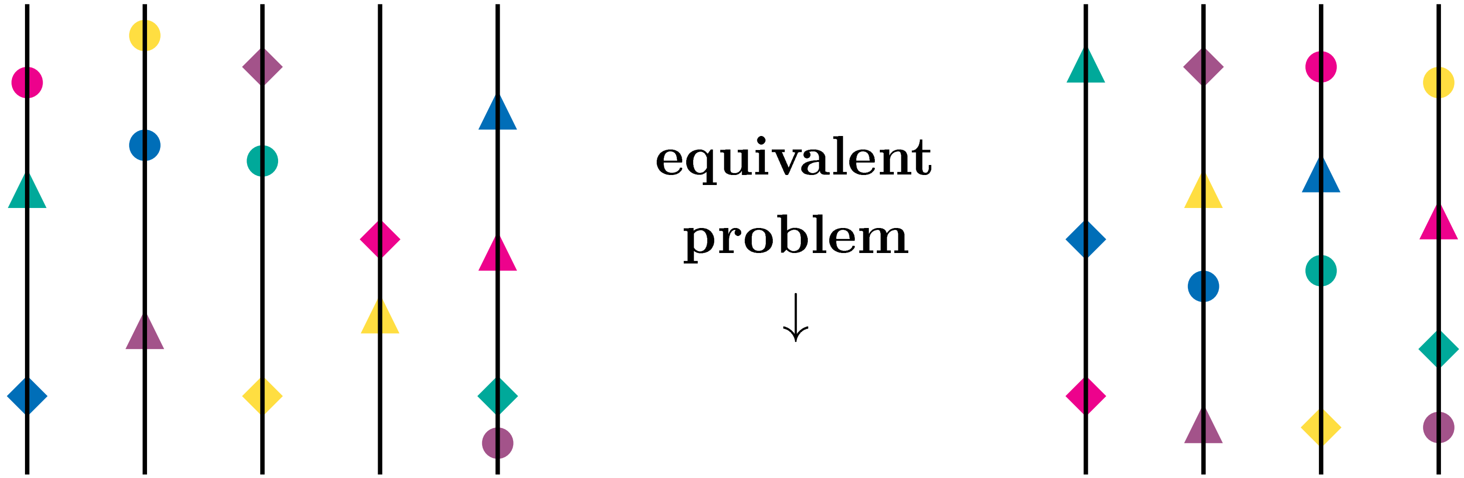
fusion

fission

translocation

Compact representation of the problem

G1 **G2**



equivalent problem

$\downarrow$

$\{1,4\}$
$\{2,3,4\}$
$\{1,3,5\}$
$\{2,5\}$

k

to be transformed into

$\longleftarrow$

with the same operations

$$\underbrace{\{1\}\{2\}\{3\}\{4\}\{5\}}_{n}$$

Compact representation of the problem

$\{1, 4\}$

$\{2, 3, 4\}$

$\{1, 3, 5\}$

$\{2, 5\}$

k

to be transformed into $\longleftarrow$

$\underbrace{\{1\}\{2\}\{3\}\{4\}\{5\}}_{n}$

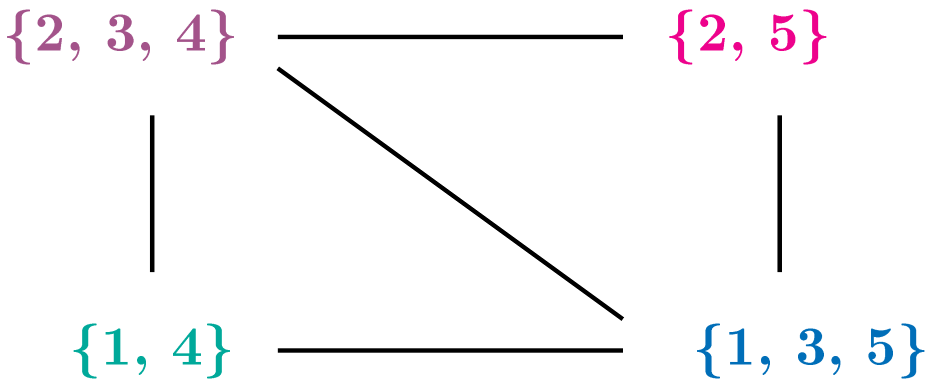
with the same operations

fusion	$(1, 3, 5)(2, 5) \rightarrow (1, 2, 3, 5)$
fission	$(1, 4) \rightarrow (1)(4)$
translocation	$(2, 3, 4)(1, 4) \rightarrow (1, 2, 3)(4)$

The graph associated with an instance of synteny

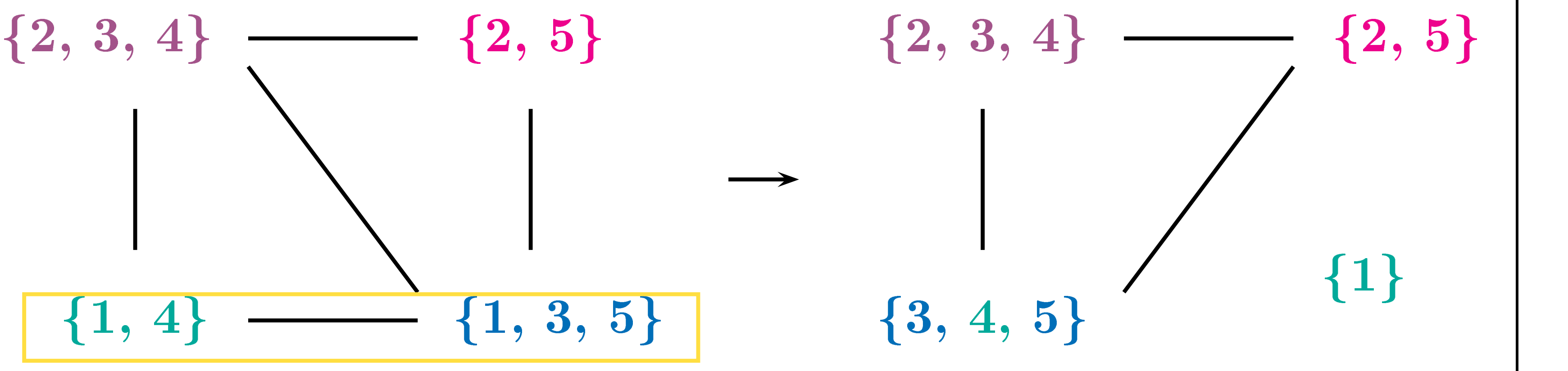
Components of an instance $\mathcal{S}(n,k)$ = components of $\mathcal{G}(V,E)$ where

- V **is the set of** k **subsets of** $\{1,\cdots,n\}$ **corresponding to the second genome in the compact representation**
- E **is the set of edges which link two vertices if the intersection of the corresponding subsets is non empty**



“Solving an element”

Performing a translocation which creates a component composed of a singleton



A lower bound for the distance (Das Gupta *et al.*)

Lower bound on the number of operations

$\mathcal{D}(\mathcal{S}(n,k)) = n - c$ **where** c **is the number of components**

Intuitive idea

Consider a single component with m **elements**

each operation can, at best, solve an element

at least $m-1$ **operations are needed to solve all** m **elements**

Minimum number of operations for instance with c **components each having** m_i **elements**

$\sum_{i=1}^{c}(m_i - 1) = (\sum_{i=1}^{c} m_i) - c = n - c$

A few further useful results (Liben-Nowell)

Monotonicity

Adding any element to any set of an instance of synteny results in a new instance that is not simpler than the original one

Canonical move sequence

For all instances $\mathcal{S}(n,k)$ there is always an optimal move sequence such that the order of the operations is

1. all fusions
2. all translocations
3. all fissions

Complexity results on syntenic distance

$\mathcal{NP}$**-hardness**	**Das Gupta, Recomb 97**
2-approximations	**Das Gupta, Recomb 97**
	Ferretti, CPM 96
	Liben-Nowell, CPM 99

$\mathcal{NP}$**-hard already for one component**

The simplest 2-approximation (hypotheses: $k \leq n,\ c = 1$**)**

- **Do** $k - 1$ **fusions**
- **Do** $n - 1$ **fissions**

Total: $n + k - 2$ **operations with** $n + k - 2 \leq 2(n-1) = 2$ **Opt.**

Addressing a couple of conjectures (due to Liben-Nowell, 1999)

Size of syntenic diameter

Feasibility of improving approximation ratio by allowing for inter-component moves

Syntenic diameter

Corresponds to the cost of the "biggest" instance of size n

Thanks to the monotonicity property, we can identify such "biggest" instance with the instance

$$\mathcal{K}(n,n) = \underbrace{\{1,2,\ldots,n\},\{1,2,\ldots,n\},\ldots,\{1,2,\ldots,n\}}_{n \text{ times}}$$

Liben-Nowell (CPM 99) conjectured that $\mathcal{K}(n,n)$ needs exactly $2n-3$ operations to be solved optimally

The syntenic diameter is no more than $2n-4$, $n \geq 4$

$$\underbrace{\{1,2,\ldots,n\},\ldots,\{1,2,\ldots,n\}}_{\text{fuse } n-4 \text{ times}} \to \underbrace{\{1,2,\ldots,n\},\ldots,\{1,2,\ldots,n\}}_{4 \text{ times}}$$

$$\text{twice } \underbrace{\{1,2,\ldots,n\},\{1,2,\ldots,n\}}_{\text{translocate}} \to \{\mathbf{1},\mathbf{2}\},\{\mathbf{3},\ldots,\mathbf{n}\}$$

$$\underbrace{\{1,2\},\{1,2\}}_{\text{translocate}} \to \{\mathbf{1}\},\{\mathbf{2}\}$$

$$\underbrace{\{3,\ldots,n\},\{3,\ldots,n\}}_{\text{translocate}} \to \{\mathbf{3}\},\{\mathbf{4},\ldots,\mathbf{n}\}$$

$$\underbrace{\{3,\ldots,n\}}_{\text{fission } n-4 \text{ times}} \to \{\mathbf{3}\},\ldots,\{\mathbf{n}\}$$

Total number of operations

$$(n-4)+2+1+1+(n-4)=2n-4$$

The syntenic diameter is exactly $2n-4$, $n \geq 4$

Intuition of the proof

- # fusions = # fissions
- If only translocations are performed, diameter is at least $2n-4$
  - no element can be solved in the first $n-2$ moves
  - after first $n-2$ moves, all elements with count 2 are in at most 4 sets and at most 2 components
  - # moves required are at least

$$(n-2)+(n_1-1)+(n_2-1)+\sum_{i=3}^{c}[(n_i-1)+1]$$

- # translocations = $\mathcal{D}(\mathcal{K}(m,m))$ where $m = n -$ #fusions
- Proof by induction for the case where fusions (and therefore fissions) are allowed

An important corollary: minimum # translocations is 4

Bidimensional syntenic diameter

Corresponds to $\mathcal{K}(n,m)$, that is, to:

$$\underbrace{\{1,2,\ldots,n\},\{1,2,\ldots,n\},\ldots,\{1,2,\ldots,n\}}_{m\ \textbf{times}}$$

This represents a more general upper bound for $\mathcal{S}(n,k)$

$\mathcal{D}(\mathcal{K}(n,m)) = m+n-4$, $\forall n,m \geq 4$

Intuition of proof

$4 \leq m \leq n$	**represents an intermediate case of $\mathcal{K}(n,n)$ with only $m-4$ fusions to do**
otherwise	**consider dual problem (same result)**

Addressing a couple of conjectures (contd.)

Size of syntenic diameter

Feasibility of improving approximation ratio by allowing for inter-component moves

Diminishing # operations by doing inter-component moves

Example

$$\underbrace{\{1, 2, \ldots, n-1\}}_{\text{once: } \mathcal{C}_1}, \qquad \underbrace{\{n\}}_{n-1 \text{ times: } \mathcal{C}_2}$$

Forbidding inter-component moves

Optimal move sequence for $\mathcal{C}_1$

Fission $n-2$ times

Optimal move sequence for $\mathcal{C}_2$

Fuse $n-2$ times

Number of operations: $2n-4$

Diminishing # operations by doing inter-component moves

Example

$$\underbrace{\{1,2,\ldots,n-1\}}_{\text{once: } \mathcal{C}_1}, \quad \underbrace{\{n\}}_{n-1 \text{ times: } \mathcal{C}_2}$$

Allowing for inter-component moves

Optimal inter-component move sequence

$$\underbrace{\{1,\ldots,n-1\},\{n\}}_{\text{translocate}} \to \{1\},\{2,\ldots,n-1\}$$

$$\underbrace{\{2,\ldots,n-1,n\},\{n\}}_{\text{translocate}} \to \{2\},\{3,\ldots,n-1\}$$

$\ldots$

$$\underbrace{\{n-1,n\},\{n\}}_{\text{translocate}} \to \{n-1\},\{n\}$$

Number of operations: $n-1$

Diminishing # operations by doing inter-component moves

A more general example

$$\underbrace{\mathcal{S} = \{x_1, x_2, \ldots, x_f\}}_{\text{once: } \mathcal{C}_1}, \qquad \underbrace{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_g}_{\text{sets which in optimal have to be fused together: } \mathcal{C}_2}$$

Forbidding inter-component moves

Optimal move sequence for $\mathcal{C}_1$

Fission $f - 1$ times

Optimal move sequence for $\mathcal{C}_2$

Requires doing $g - 1$ fusions

Number of operations: $f + g - 2$

Diminishing # operations by doing inter-component moves

Same example

$$\underbrace{\mathcal{S} = \{x_1, x_2, \ldots, x_f\}}_{\text{once: } \mathcal{C}_1}, \qquad \underbrace{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_g}_{\text{sets which in optimal have to be fused together: } \mathcal{C}_2}$$

Allowing for inter-component moves

Optimal inter-component move sequence ($f \leq g$)

$$\underbrace{\{x_1, \ldots, x_f\}, \mathcal{S}_1}_{\text{translocate}} \to \{x_1\}, \{x_2, \ldots, x_f\} \cup \mathcal{S}_1$$

$$\underbrace{\{x_2, \ldots, x_f\} \cup \mathcal{S}_1, \mathcal{S}_2}_{\text{translocate}} \to \{x_2\}, \{x_3, \ldots, x_f\} \cup \mathcal{S}_1 \cup \mathcal{S}_2$$

$\ldots$

$$\underbrace{\{x_f\} \cup \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_{f-1}, \mathcal{S}_f}_{\text{translocate}} \to \{x_f\}, \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_f$$

Number of operations spared: $\min\{f, g\} - 2$

“Optimal component pairing”

Let $\mathcal{S}(n,k)$ be an instance composed of:

- **$c \geq 2$ components $\mathcal{C}_1, \cdots, \mathcal{C}_c$ where each $\mathcal{C}_i$ contains g_i sets which, in an optimal move sequence, need to be fused**
- **$d \geq 1$ components $\mathcal{C}'_1, \cdots, \mathcal{C}'_d$ with $\mathcal{C}'_j = \{x^j_1, \cdots, x^j_{f_j}\}$**

If inter-component moves are allowed, then an optimal move sequence solving $\mathcal{S}(n,k)$ will be one which, *at the same time*:

- **performs the maximum number of pairings between components $\mathcal{C}_1, \cdots, \mathcal{C}_c$ and components $\mathcal{C}'_1, \cdots, \mathcal{C}'_d$**
- **translocates each component $\mathcal{C}_i$ for $1 \leq i \leq c$ with a minimum number of different components among $\mathcal{C}'_1, \cdots, \mathcal{C}'_d$**

Alas, finding an optimal component pairing is $\mathcal{NP}$-hard

Reduction to the 3-Partition problem (sketch)

Reminder

input

a finite set A **of** $3m$ **elements**

a bound $B \in \mathbb{Z}^+$

a size $s(a) \in \mathbb{Z}^+$ **for each** $a \in A$ **such that**

$B/4 < s(a) < B/2$ **and** $\sum_{a \in A} s(a) = mB$

output

can A **be partitioned into** m **disjoint sets**

$A_1, A_2, \ldots, A_m$ **such that** $\sum_{a \in A_i} s(a) = B$ **?**

Alas, finding an optimal component pairing is $\mathcal{NP}$-hard

Reduction to the 3-Partition problem (sketch)

Reduction: sketch of proof

Consider an arbitrary instance of the 3-Partition problem. We build an instance:

$$\mathcal{S}(n = 3m + Bm, k = (3+1)m)$$

in the following way. The instance will contain:

- $3m$ **components** $\mathcal{C}_1, \ldots, \mathcal{C}_{3m}$ **having each** $s(C_i)$ **copies of a same set** $\mathcal{S}_i = \{y_i\}$ **with** $B/4 < s(C_i) < B/2$ **for a** $B \in \mathbb{Z}^+$
- m **distinct components** $\mathcal{C}'_1, \ldots, \mathcal{C}'_m$ **each containing a unique set with** B **distinct elements** $\{x^i_1, \ldots, x^i_B\}$ **which need to be fissioned**

Alas, finding an optimal component pairing is $\mathcal{NP}$-hard

Reduction to the 3-partition problem (contd.)

We must find a pairing which makes the maximum number of inter-component translocations yet which, at the same time, translocates each of $\mathcal{C}_1, \ldots, \mathcal{C}_{3m}$ with the minimum number of $\mathcal{C}'_j$ components for $1 \leq j \leq m$

Since the number $s(C_i)$ of elements in each component $\mathcal{C}_i$ is such that $B/4 < s(C_i) < B/2$, the minimum will be reached if we can translocate exactly 3 components among $\mathcal{C}_1, \ldots, \mathcal{C}_{3m}$ with each component $\mathcal{C}'_1, \cdots, \mathcal{C}'_m$

Addressing a couple of conjectures (due to Liben-Nowell, 1999)

Size of syntenic diameter

Solved (also, independently, by Liben-Nowell and Kleinberg)

Feasibility of improving approximation ratio by allowing for inter-component moves

Still an open problem !

Constrained versions of synteny

Constrained move sequence

Constrained instance

Constrained move sequence: linear synteny

Linear synteny $\mathcal{L}(n,k)$

The optimal linear move sequence must obey the following constraints

- **the first $k-1$ moves must be either fusions or translocations solving one element**
  - **translocations solving no element are forbidden**
- **the remaining are fissions**

Why is it interesting ?

$\mathcal{D}(\mathcal{S}(\mathbf{n},\mathbf{k})) \leq \mathcal{D}(\mathcal{L}(\mathbf{n},\mathbf{k})) \leq \mathcal{D}(\mathcal{S}(\mathbf{n},\mathbf{k})) + \log_{\frac{4}{3}} \mathcal{D}(\mathcal{S}(\mathbf{n},\mathbf{k}))$
(DasGupta *et al.***)**

Constrained instance: exact and nested synteny

Exact synteny

An instance $\mathcal{S}(n,k)$ such that $\mathcal{D}(\mathcal{S}(n,k)) = n-1$

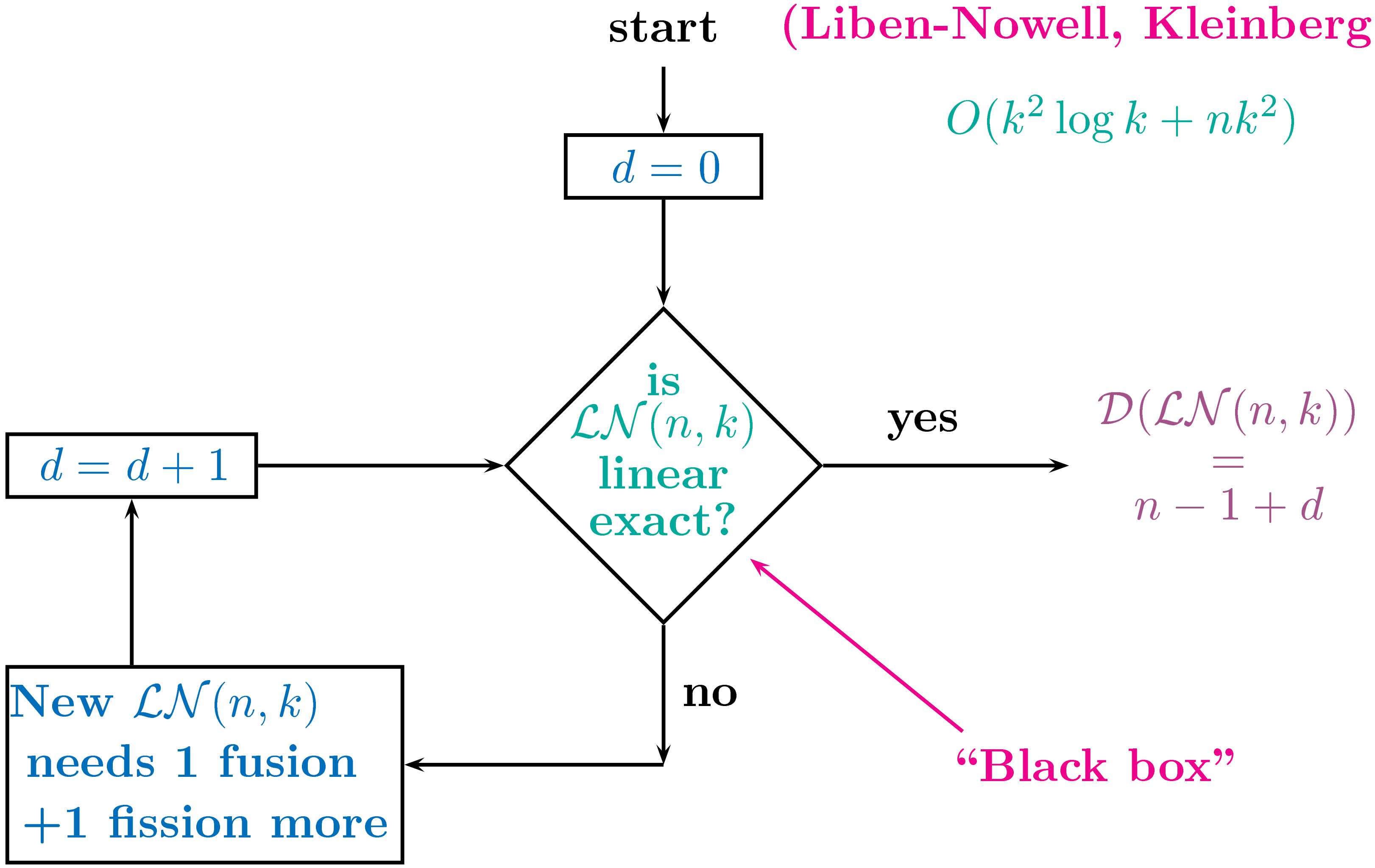
Nested synteny

An instance $\mathcal{S}(n,k)$ such that, $\forall \mathcal{S}_i, \mathcal{S}_j \in \mathcal{S}(n,k)$, either $\mathcal{S}_i \subseteq \mathcal{S}_j$ or $\mathcal{S}_j \subseteq \mathcal{S}_i$

A polynomial algorithm for linear nested synteny

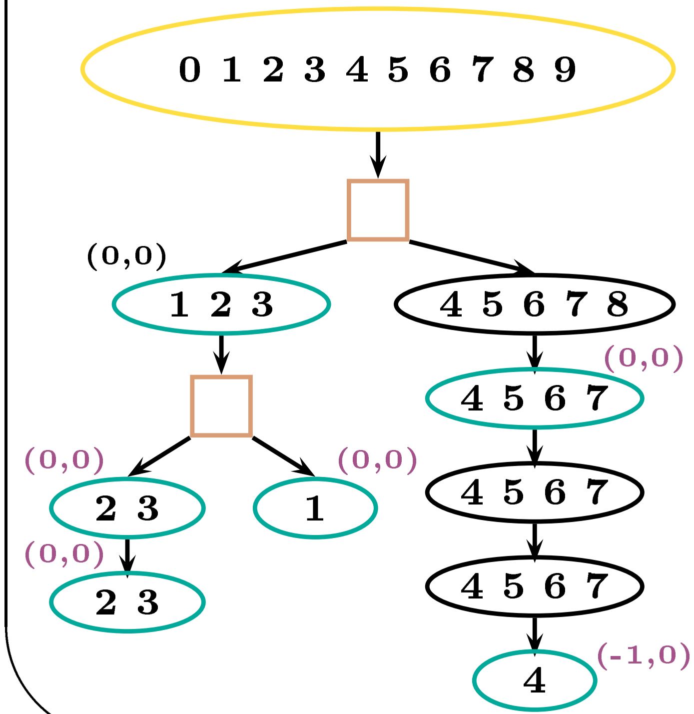
(Liben-Nowell, Kleinberg 99)

$O(k^2 \log k + nk^2)$



A simpler black box for linear nested synteny

It answers the question "Is $\mathcal{LN}(n,k)$ linear exact ?" in $O(k^2 + nk + k\log k)$ time by using a graph representation of the instance



dummy node

frontier node

(v_b,v_t) at frontier node:

v_b: # moves solving no element

v_b: # elements left in single copy

General nested synteny

An example

$\{1,2,3,4\},\{1,2,3,4\},\{1,2\},\{3,4\}$

A linear optimal move sequence

- **Since we cannot solve any element, the first operation must be a fusion**
  - **the best one is of set $\{1,2,3,4\}$ with the other set $\{1,2,3,4\}$**
- **We then translocate $\{1,2,3,4\}$ with $\{1,2\}$ solving 1 (for instance)**
- **We follow by a translocation of $\{2,3,4\}$ with $\{3,4\}$ solving 3 (for instance)**
- **We end up by fissioning $\{2,4\}$**

Total: 4 moves

General nested synteny (contd.)

Same example

$\{1,2,3,4\},\{1,2,3,4\},\{1,2\},\{3,4\}$

An optimal move sequence

We start with one translocation between $\{1,2,3,4\}$ and $\{1,2,3,4\}$ which solves no element but allows to partition the sets and obtain $\{1,2\}$ and $\{3,4\}$

We translocate $\{1,2\}$ with $\{1,2\}$ solving both 1 and 2

We translocate $\{3,4\}$ with $\{3,4\}$ solving both 3 and 4

Total: 3 moves

What I shall not have time to talk about

A polynomial solution for general nested synteny that is conjectured to be optimal and uses:

- **This idea of partitioning the bigger sets according to the "children" they include**
- **The black box just shown to check whether some of the parents+children thus obtained are "exact" instances**

An inclusion theorem for the general synteny instance $\mathcal{S}(n,k) = \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ which states that there always exists an optimal move sequence where $\mathcal{S}_j$ is used before $\mathcal{S}_i$ if $\mathcal{S}_i \subset \mathcal{S}_j$

A few perspectives

Find a better than 2-approximation algorithm (in the case of a multi-component instance, by allowing for moves between components)

Prove the conjecture of the previous slide

(The main one) Treat other rearrangement distances