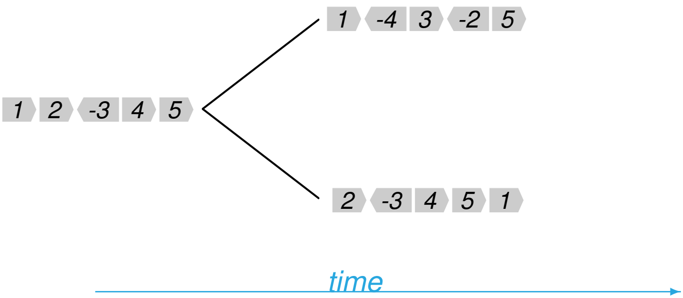
Universität Bielefeld
Technische Fakultät
Genominformatik

Phylogeny-based Analysis of Gene Clusters

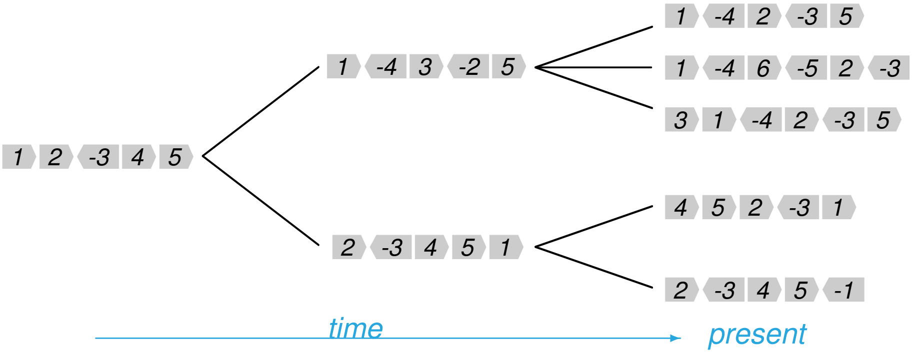
Roland Wittler

Introduction

Introduction



Introduction



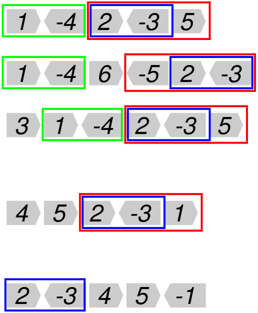
Introduction

Whole-Genome Comparison

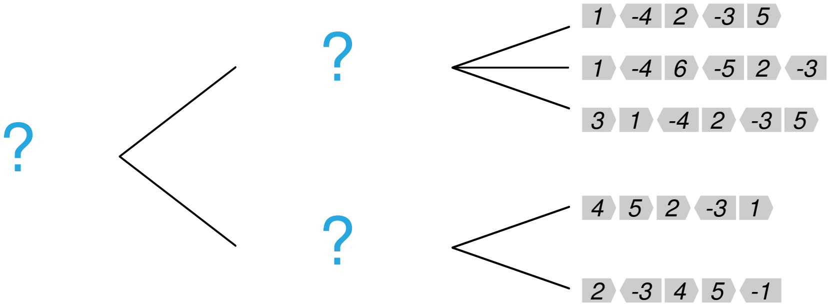
gene cluster

$\approx$ co-located genes

$\leftrightsquigarrow$ functionally related



Introduction



?

?

1 -4 2 -3 5

1 -4 6 -5 2 -3

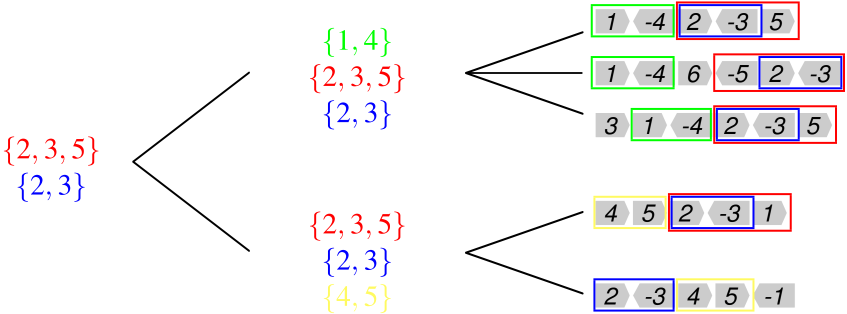
3 1 -4 2 -3 5

?

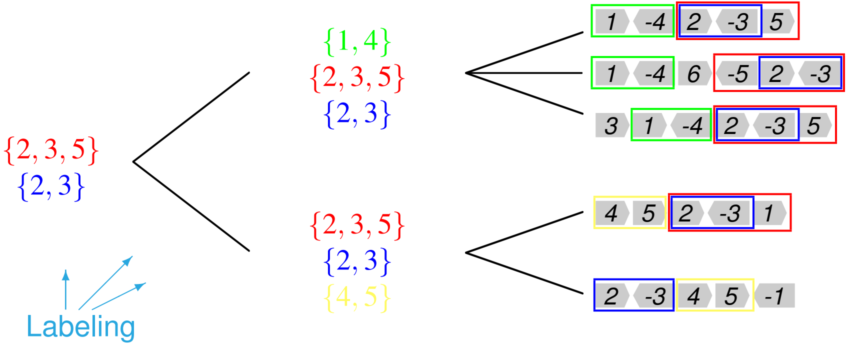
4 5 2 -3 1

2 -3 4 5 -1

Introduction



Introduction



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

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

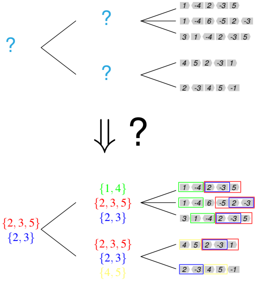
$\{2,3,5\}$
$\{2,3\}$
$\{4,5\}$

Labeling

Introduction

Goals

- objective function
- exact method
- different cluster models



Labeling Problem

Models for genomes

permutations

sequences

multichromosomal

circular

...

Models for gene clusters

common intervals

max-gap cluster

conserved intervals

adjacencies

...

Labeling Problem

Abstract gene cluster model

How do genomes look like?

Labeling Problem

Abstract gene cluster model

$\mathcal{U}$ – the universal set, i.e. the set of all possible genomes

Labeling Problem

Abstract gene cluster model

$\mathcal{U}$ – the universal set, i.e. the set of all possible genomes

How do gene clusters look like?

Labeling Problem

Abstract gene cluster model

$\mathcal{U}$ – the universal set, i.e. the set of all possible genomes

$\mathcal{C}$ – the set of all possible gene clusters

Labeling Problem

Abstract gene cluster model

$\mathcal{U}$ – the universal set, i.e. the set of all possible genomes

$\mathcal{C}$ – the set of all possible gene clusters

When is a cluster contained in a genome?

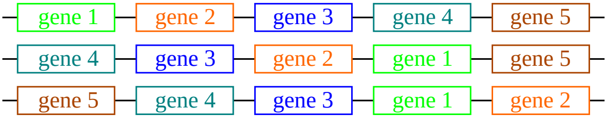
Labeling Problem

Abstract gene cluster model

$\mathcal{U}$ –	the universal set, i.e. the set of all possible genomes
$\mathcal{C}$ –	the set of all possible gene clusters
$\sqsubseteq$ –	determines whether a gene cluster $c \in \mathcal{C}$ is contained in a genome $g \in \mathcal{U}$

Labeling Problem

Example: Adjacencies on permutations



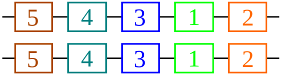
$\mathcal{U} := \{$

— gene 1 — gene 2 — gene 3 — gene 4 — gene 5 — ,

— gene 4 — gene 3 — gene 2 — gene 1 — gene 5 — ,

— gene 5 — gene 4 — gene 3 — gene 1 — gene 2 — , … $\}$

$\mathcal{C} := \{ \; \{1,2\}, \{1,3\}, \{2,3\}, \{2,5\}, \ldots \; \}$



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

$\{2,3\} \not\sqsubseteq$ — 5 — 4 — 3 — 1 — 2 —

Example: Common Intervals on permutations

$$\mathcal{U} := \left\{ \begin{array}{l} \boxed{\text{gene 1}}\boxed{\text{gene 2}}\boxed{\text{gene 3}}\boxed{\text{gene 4}}\boxed{\text{gene 5}}\,, \\ \boxed{\text{gene 4}}\boxed{\text{gene 3}}\boxed{\text{gene 2}}\boxed{\text{gene 1}}\boxed{\text{gene 5}}\,, \\ \boxed{\text{gene 5}}\boxed{\text{gene 4}}\boxed{\text{gene 3}}\boxed{\text{gene 1}}\boxed{\text{gene 2}}\,, \ldots \end{array} \right\}$$

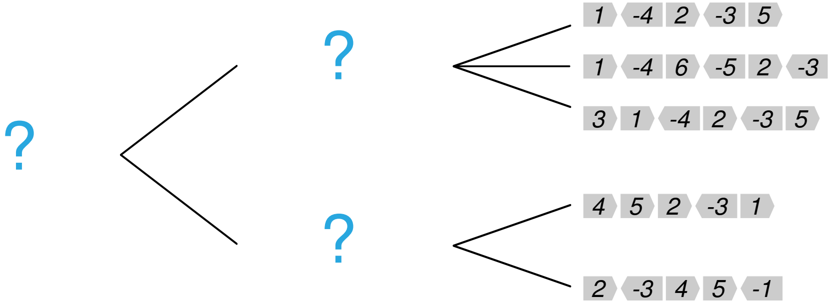
$$\mathcal{C} := \{ \quad \{1,2\}\,,\{1,3\}\,,\{1,2,3\}\,,\{2,3,5\}\,,\ldots \quad \}$$

$$\{1,2,3\} \quad ⋿ \quad \boxed{5}\boxed{4}\boxed{3}\boxed{1}\boxed{2}$$

$$\{2,3,5\} \quad \not⋿ \quad \boxed{5}\boxed{4}\boxed{3}\boxed{1}\boxed{2}$$

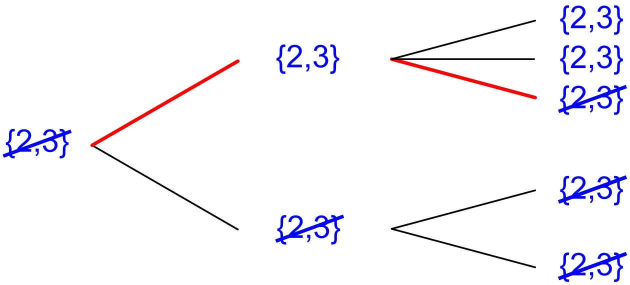
Labeling Problem

Objective: Parsimony & Consistency



Labeling Problem

Parsimony

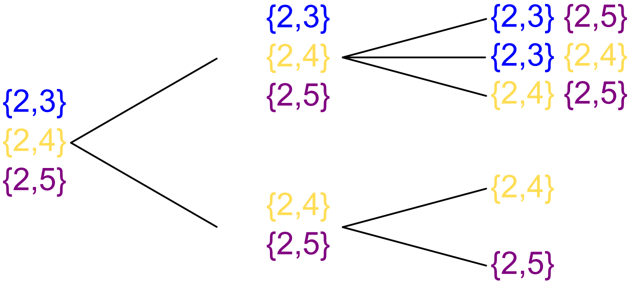


$weight = \#\text{gains} + \#\text{losses} = 2$

minimal ⇒ parsimonious

Labeling Problem

Consistency



{2,3}
{2,4}
{2,5}

{2,3}
{2,4}
{2,5}

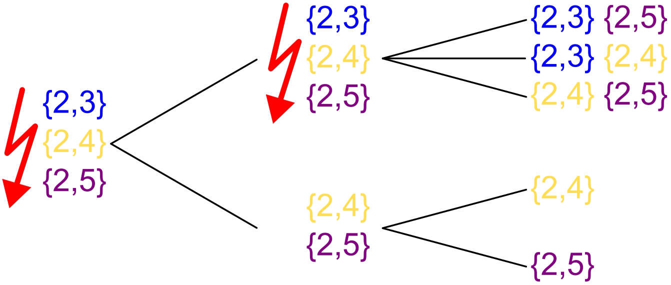
{2,3} {2,5}
{2,3} {2,4}
{2,4} {2,5}

{2,4}
{2,5}

{2,4}
{2,5}

Labeling Problem

Consistency



{2,3}
{2,4}
{2,5}

{2,3}
{2,4}
{2,5}

{2,3} {2,5}
{2,3} {2,4}
{2,4} {2,5}

{2,4}
{2,5}

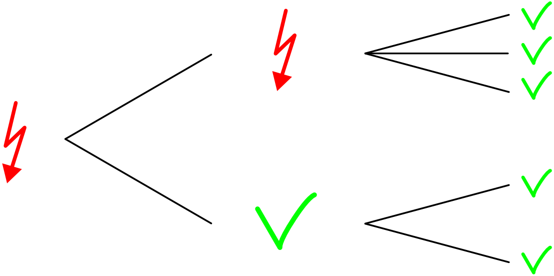
{2,4}

{2,5}

inconsistent

Labeling Problem

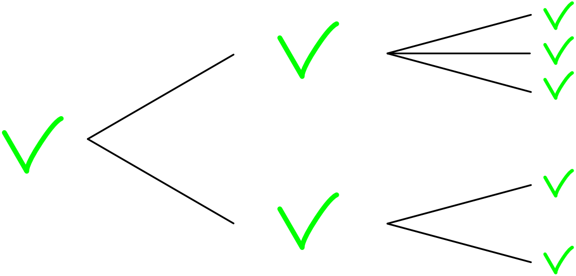
Consistency



inconsistent

Labeling Problem

Consistency



consistent

Labeling Problem

Overall Objective

$$\left.\begin{array}{l}\text{Under all consistent labelings,}\\ \text{find a most parsimonious one.}\end{array}\right\}\ \text{optimal}$$

Algorithmical Framework

General approach (based on abstract cluster model)

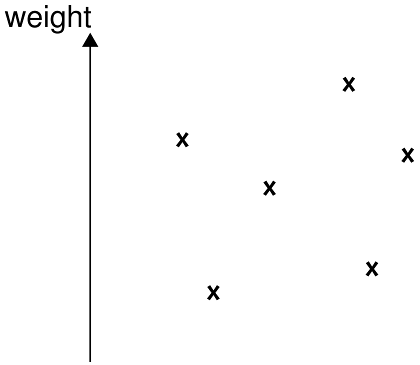
1. Initialization: Parsimony
2. delete conflicting clusters

⇒ Consistency

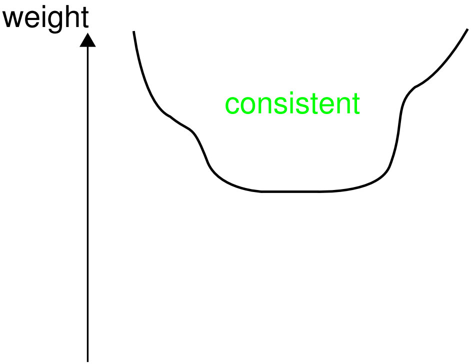
$\left.\begin{array}{l}\\ \\ \\ \end{array}\right\}$ Optimality

Thereom: Exactness ✓

Searchspace

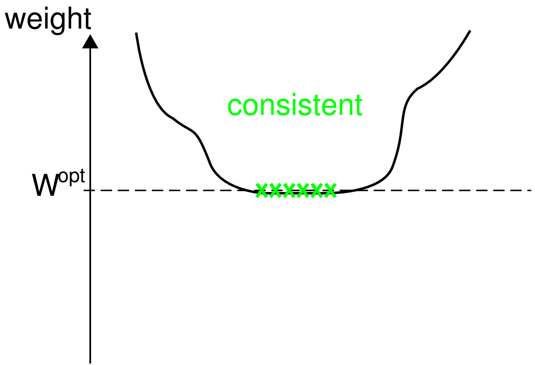


Searchspace



Searchspace

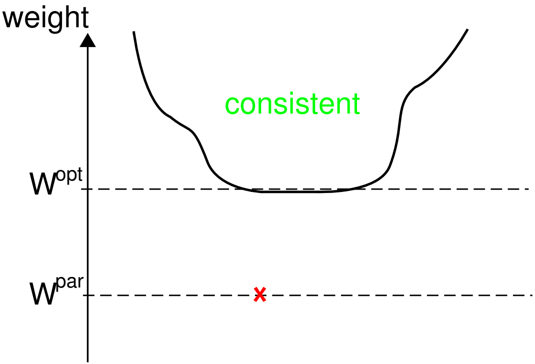
Optimality



Under all consistent labelings, find a most parsimonious one.

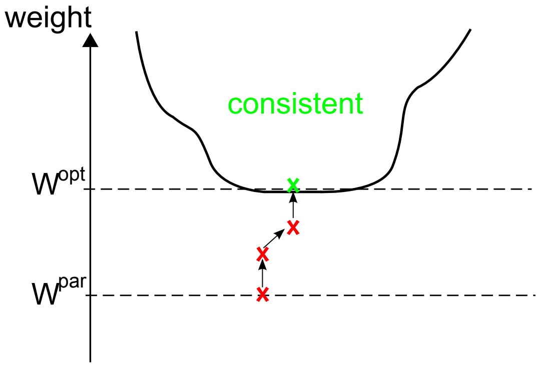
Searchspace

Theorem



Searchspace

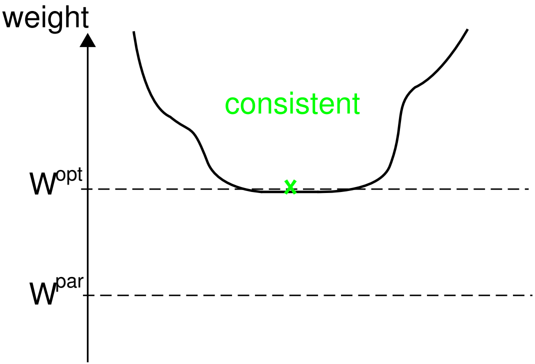
Theorem



- deletion:
  $\{2,3\} \longrightarrow \cancel{\{2,3\}}$

Searchspace

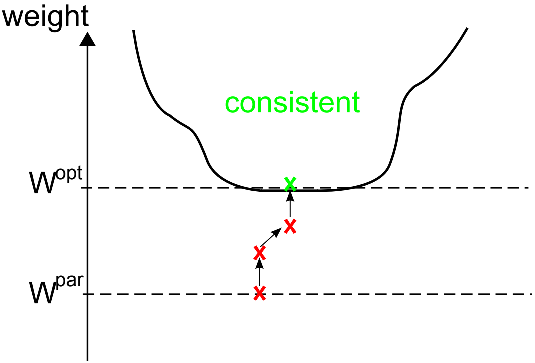
Theorem



- deletion:
  $\{2,3\} \longrightarrow \cancel{\{2,3\}}$
- two directions

Searchspace

Theorem



- deletion:
  {2,3} ⟶ ~~{2,3}~~
- two directions

Algorithmical Framework

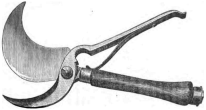
Branch & Bound Approach

Branch

- for each node:
  for each cluster:
  delete, recurse

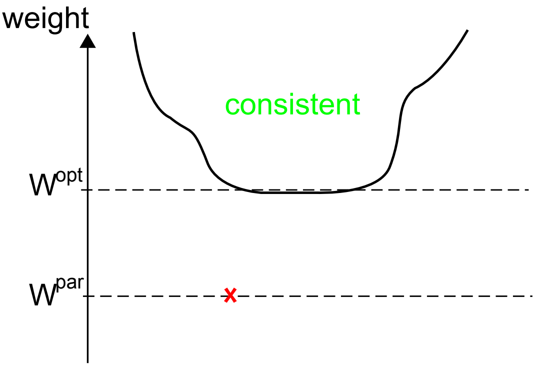
Bound

- keep track of current optimum



Searchspace

Branch and Bound

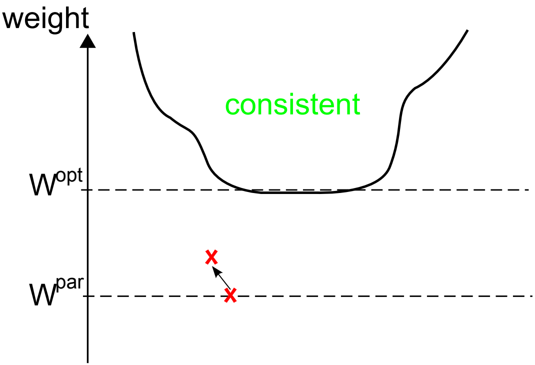


Branch

```
for each node:
  for each cluster:
    delete, recurse
```

Searchspace

Branch and Bound

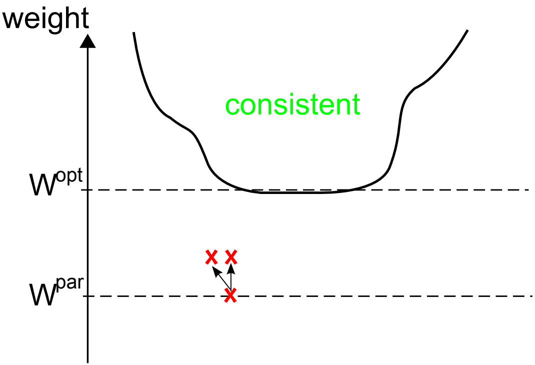


Branch

```
for each node:
  for each cluster:
    delete, recurse
```

Searchspace

Branch and Bound

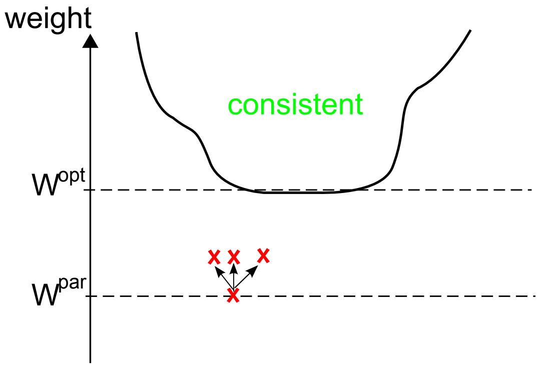


Branch

```
for each node:
  for each cluster:
    delete, recurse
```

Searchspace

Branch and Bound

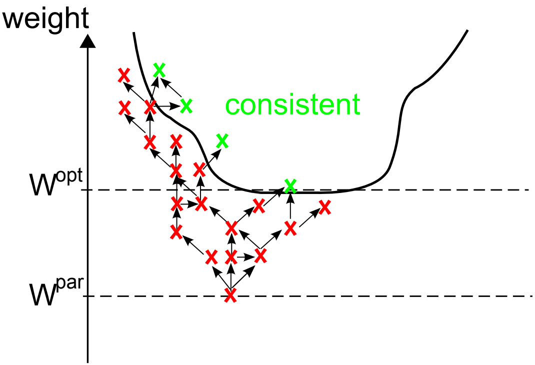


Branch

```
for each node:
  for each cluster:
    delete, recurse
```

Searchspace

Branch and Bound

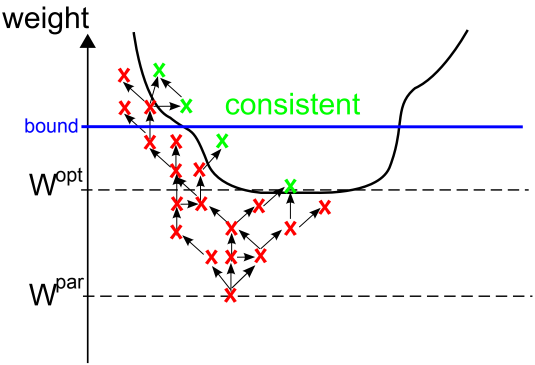


Branch

```
for each node:
  for each cluster:
    delete, recurse
```

Searchspace

Branch and Bound



Branch

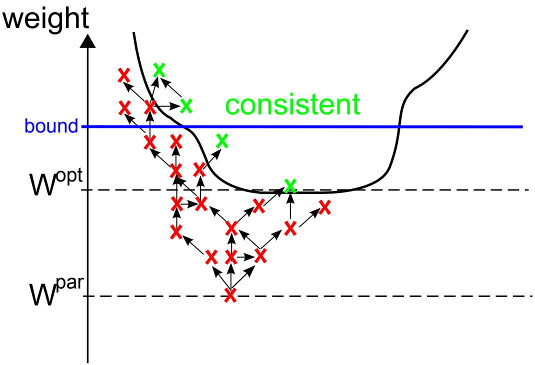
for each node:
  for each cluster:
    delete, recurse

Bound

keep track of
current optimum

Searchspace

Branch and Bound



Branch

for each node:
  for each cluster:
    delete, recurse

Bound

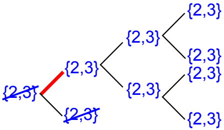
keep track of
current optimum

Monotonicity

preclude decrease
of weight

Searchspace

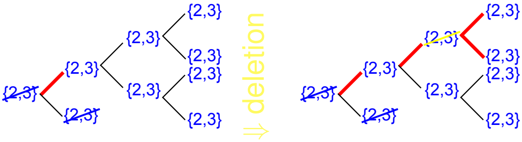
Reoptimization



weight $= 1$

Searchspace

Reoptimization

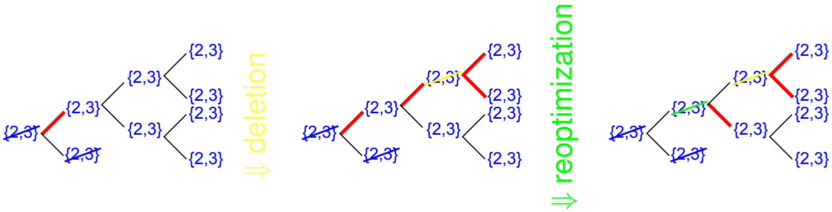


weight $= 1$

weight $= 4$

Searchspace

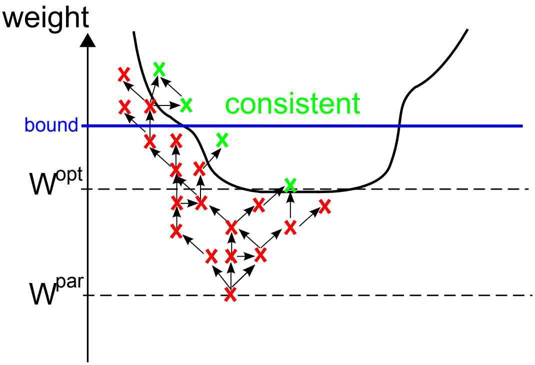
Reoptimization



weight $= 1$ weight $= 4$ weight $= 3$

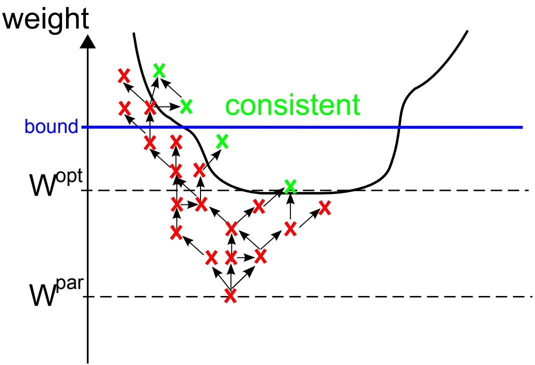
Searchspace

Branch and Bound



Searchspace

Branch and Bound — Speed-up



Branch

in promising order:

most conflicting first

estimate CI

Greedy — Consistency revisited

{2,3}
{2,4}
{2,5}

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g$ contains all clusters in $C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \setminus \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\}$

$\mathrm{CI}(C,\{3,4\}) = 2$

$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

{2,3}

{2,4}

{2,5}

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \backslash \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\}$

$\mathrm{CI}(C,\{3,4\}) = 2$

$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \backslash \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\}$

$\mathrm{CI}(C,\{3,4\}) = 2$

$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$
$C = \big\{\{3,4\},\{4,5\}\big\}$
$\mathrm{GS}(C) = \big\{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\big\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \backslash \{c\}) \neq \emptyset$
$C_1 = \big\{\{3,4\},\{4,5\},\{4,6\}\big\}$
$C_2 = \big\{\{3,4\},\{4,5\},\{5,3\}\big\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c
$C = \big\{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\big\}$
$\mathrm{CI}(C,\{3,4\}) = 2$
$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \setminus \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\}$

$\mathrm{CI}(C,\{3,4\}) = 2$

$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \big\{\{3,4\},\{4,5\}\big\}$

$\mathrm{GS}(C) = \big\{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\big\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C$: $\mathrm{GS}(C \backslash \{c\}) \neq \emptyset$

$C_1 = \big\{\{3,4\},\{4,5\},\{4,6\}\big\}$

$C_2 = \big\{\{3,4\},\{4,5\},\{5,3\}\big\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \big\{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\big\}$

$\mathrm{CI}(C,\{3,4\}) = 2$

$\mathrm{CI}(C,\{5,3\}) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \backslash \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\}$

$\mathrm{CI}\big(C,\{3,4\}\big) = 2$

$\mathrm{CI}\big(C,\{5,3\}\big) = 1$

Greedy — Consistency revisited

genome set $\mathrm{GS}(C) := \{g \in \mathcal{U} \mid g \text{ contains all clusters in } C\}$

$C = \{\{3,4\},\{4,5\}\}$

$\mathrm{GS}(C) = \{(\ldots,3,4,5,\ldots),(\ldots,5,4,3,\ldots)\}$

conflicting set $\mathrm{GS}(C) = \emptyset$ and for all $c \in C : \mathrm{GS}(C \backslash \{c\}) \neq \emptyset$

$C_1 = \{\{3,4\},\{4,5\},\{4,6\}\}$

$C_2 = \{\{3,4\},\{4,5\},\{5,3\}\}$

conflict index $\mathrm{CI}(C,c) := \#$ confl. subsets of C that contain c

$C = \big\{\{3,4\},\{4,5\},\{4,6\},\{5,3\}\big\}$

$\mathrm{CI}\big(C,\{3,4\}\big) = 2$

$\mathrm{CI}\big(C,\{5,3\}\big) = 1$

Algorithmical Framework

Two-Phase Approach

sparse approach choose $\cancel{\{2,3\}}$ whenever possible

$\Rightarrow$ few clusters

$\Rightarrow$ few conflicts

$\Rightarrow$ fast branch and bound search

Algorithmical Framework

Two-Phase Approach

sparse approach choose ~~{2,3}~~ whenever possible
⇒ few clusters
⇒ few conflicts
⇒ fast branch and bound search

dense approach choose {2,3} whenever possible
⇒ more clusters
⇒ more conflicts
⇒ longer branch and bound search

Algorithmical Framework

Two-Phase Approach

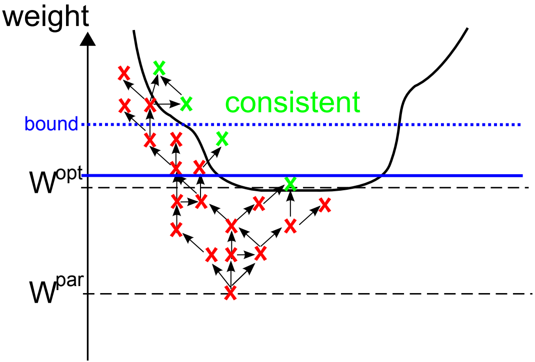
sparse approach choose ~~{2,3}~~ whenever possible
⇒ few clusters
⇒ few conflicts
⇒ fast branch and bound search

dense approach choose {2,3} whenever possible
⇒ more clusters
⇒ more conflicts
⇒ longer branch and bound search

but can be bounded by sparse optimum

Searchspace

Branch and Bound — Speed-up



Branch

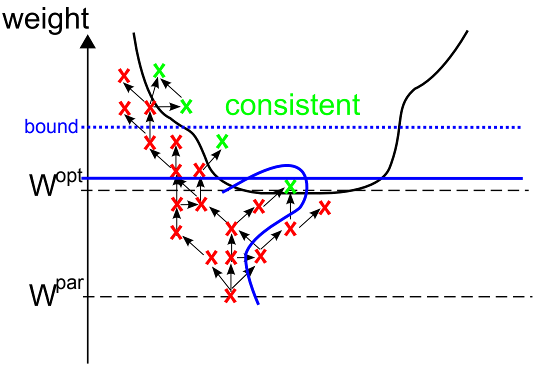
- in promising order:
  - most conflicting first
- estimate CI

Bound

- use of known
- optimal weight

Searchspace

Branch and Bound — Speed-up



Branch
in promising order:
most conflicting first
estimate CI

Bound
use of known
optimal weight

Algorithmical Framework

Complexity

worst case	exponential in $\lvert\mathcal{C}\rvert$ (number of clusters) exponential in $\lvert V\rvert$ (number of nodes)
best case	no conflicts $\Rightarrow$ only Fitch/Hartigan: $\mathcal{O}\left(\lvert V\rvert\lvert\mathcal{C}\rvert\right)$

Algorithmical Framework

Complexity

worst case exponential in $|\mathcal{C}|$ (number of clusters)
exponential in $|V|$ (number of nodes)

best case no conflicts $\Rightarrow$ only Fitch/Hartigan: $\mathcal{O}\left(|V||\mathcal{C}|\right)$

practice parsimonious weight $\approx$ optimal weight
$\Rightarrow$ sparse phase $\approx$ best case
$\Rightarrow$ low bound for dense phase

Consistency

- recursive algorithm
  to identify all conflicts
- based on abstract cluster model ($\mathbb{GS}$)
- exponential in
  - $|C|$ (number of clusters)
  - m (number of conflicts)

Search for Conflicts

Locate a subset which contains a conflicting subset

$\{ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet}_{\not\rightarrow \emptyset} \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \qquad \bullet}_{\not\to \emptyset} \qquad \circ \qquad \circ \qquad \circ \qquad \circ \qquad \circ \qquad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet}_{\not\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet}_{\not\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\circ \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\not\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \circ \quad \bullet \quad \bullet \quad \bullet}_{\not\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \circ \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \circ \quad \circ}_{\not\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\rightarrow \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \circ \quad \bullet}_{\text{conflicting subset}} \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \circ \quad \bullet}_{\text{conflicting subset}} \quad \circ \quad \circ \quad \circ \quad \}$$

Recursion

$$\{ \quad \emptyset \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$
$$\{ \quad \circ \quad \emptyset \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$
$$\{ \quad \circ \quad \circ \quad \circ \quad \circ \quad \emptyset \quad \circ \quad \circ \quad \circ \quad \}$$

Search for Conflicts

Locate a subset which contains a conflicting subset

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet}_{\to \emptyset} \quad \circ \quad \circ \quad \circ \quad \}$$

Shrink it

$$\{ \quad \underbrace{\bullet \quad \bullet \quad \circ \quad \circ \quad \bullet}_{\text{conflicting subset}} \quad \circ \quad \circ \quad \circ \quad \}$$

Recursion

$$\{ \quad \not\circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$
$$\{ \quad \circ \quad \not\circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \circ \quad \}$$
$$\{ \quad \circ \quad \circ \quad \circ \quad \circ \quad \not\circ \quad \circ \quad \circ \quad \circ \quad \}$$

- Polynomial with resp. to $|I(G)|$
- Exponential with resp. to the number of conflicting subsets

Consistency Problem

Consistency

Given

- gene cluster model
- set of clusters

{2,3}
{2,4}
{2,5}

Question

- Does a valid genome exist?

↯ ? ✓

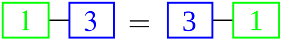
Consistency Problem

Models cluster genome	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup’s)					
sequences (incl. dup’s)					

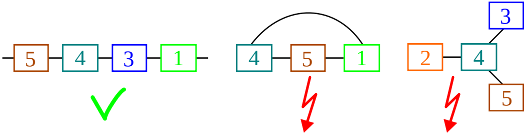
Consistency Problem

Unsigned Adjacencies on permutations

Dandekar et al., 1998



⇒ graph representation



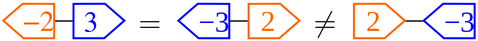
Consistency Problem

Models / cluster / genome	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph				
sequences (incl. dup's)					

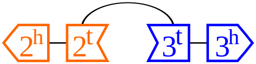
Consistency Problem

Signed Adjacencies on permutations

Dandekar et al., 1998



⇒ graph representation



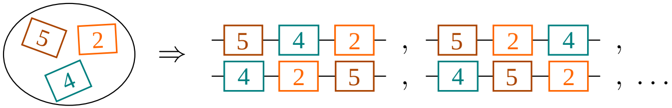
Consistency Problem

Models genome / cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph			
sequences (incl. dup's)					

Consistency Problem

Common Intervals on permutations

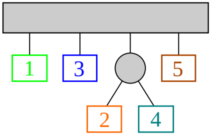
Uno & Yagiura, 2000



Consecutive Ones Property ⇒ PQ trees

Landau et al., 2005

$$\begin{array}{ccccc} \boxed{1} & \boxed{2} & \boxed{3} & \boxed{4} & \boxed{5} \\ \end{array}$$
$$\left(\begin{array}{ccccc} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{array}\right)$$



Booth & Lueker, 1976

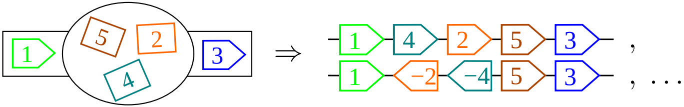
Consistency Problem

Models genome cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees		
sequences (incl. dup's)					

Consistency Problem

Framed Common Intervals on permutations

Bergeron & Stoye, 2003



≈ Common Intervals ⇒ PQ trees

Bergeron et al., 2004

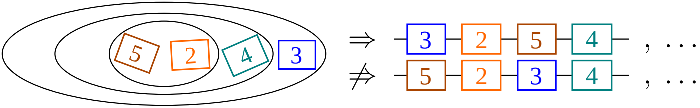
Consistency Problem

Models genome cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	
sequences (incl. dup's)					

Consistency Problem

Nested Common Intervals on permutations

Hoberman & Durand, 2005



⇒ common intervals

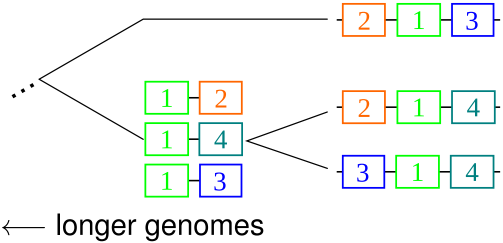
Consistency Problem

Models cluster genome	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)					

Consistency Problem

Allowing duplicated genes

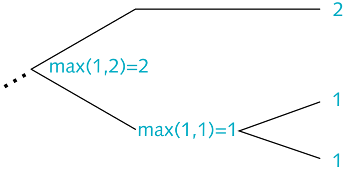
- No limitation of gene occurrences



Consistency Problem

Allowing duplicated genes

- Limit multiplicity of genes



lineage sorting, prior knowledge, parsimony, ...

Consistency Problem

Given

- gene cluster model
- set of clusters
- multiplicity $m(g)$ for each gene g

{2,3}
{2,4}
{2,5}

Question

- Does a valid genome exist ...
- ... that contains each gene g at most $m(g)$ times?

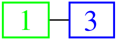
↯ ? ✓

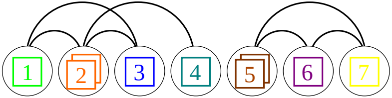
Consistency Problem

Models / cluster / genome	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)					

Consistency Problem

Unsigned Adjacencies on sequences

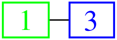


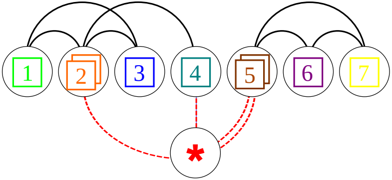


- for all genes:
  $\deg(v_g) \leq 2m(g)$
- for at least one gene in each component:
  $\deg(v_g) < 2m(g)$

Consistency Problem

Unsigned Adjacencies on sequences



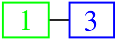


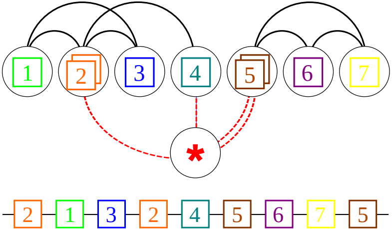
- for all genes:
  $\deg(v_g) \leq 2m(g)$
- for at least one gene in each component:
  $\deg(v_g) < 2m(g)$

⇔ graph is Eulerian

Consistency Problem

Unsigned Adjacencies on sequences





- for all genes:
  $\deg(v_g) \leq 2m(g)$
- for at least one gene in each component:
  $\deg(v_g) < 2m(g)$

$\Leftrightarrow$ graph is Eulerian

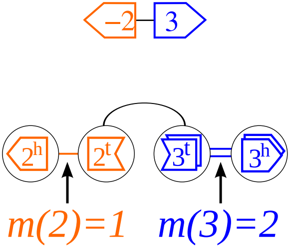
$\Leftrightarrow$ consistency

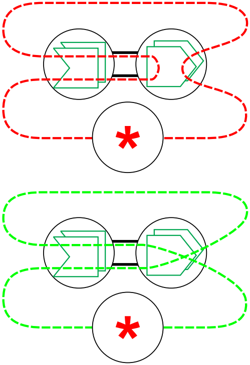
Consistency Problem

Models cluster genome	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)	linear Euler				

Consistency Problem

Signed Adjacencies on sequences



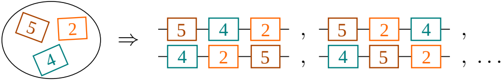


Consistency Problem

Models / genome / cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)	linear Euler	linear Euler			

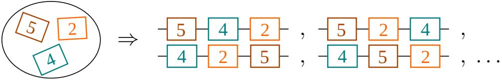
Consistency on Sequences

Common Intervals on sequences



Consistency on Sequences

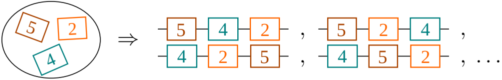
Common Intervals on sequences



- NP-complete, even if
  - maximum cluster size = 5

Consistency on Sequences

Common Intervals on sequences



- NP-complete, even if
  - maximum cluster size = 3
  - maximum multiplicity = 2
  - sum of multiplicities given

Ján Manuch & Murray Patterson

NP-Completeness

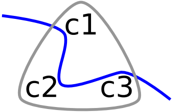
Reduction from 3SAT3

- 2-clauses and 3-clauses
- each variable: 2x positive, 1x negative

3-clause $(c1 \vee c2 \vee c3)$

literal in the middle

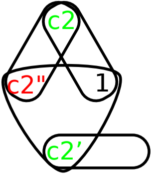
$\Rightarrow$ TRUE



NP-Completeness

Variable

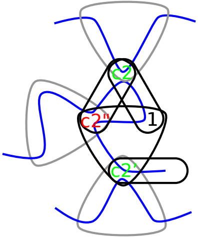
- pos.: c2, c2’
- neg.: c2”
- additional objects with $m = 1$



NP-Completeness

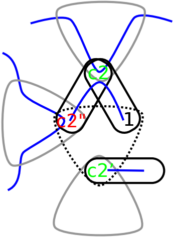
Valid:

c2=c2’=TRUE, c2”=FALSE



Invalid:

c2=TRUE, c2”=TRUE

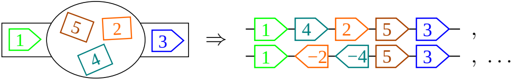


Consistency on Sequences

Models genome cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)	linear Euler	linear Euler	NP-comp. 3SAT3		

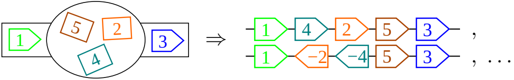
Consistency on Sequences

Framed Common Intervals on sequences

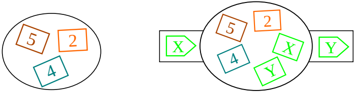


Consistency on Sequences

Framed Common Intervals on sequences



- NP-complete
- follows from NP-completeness for common intervals...

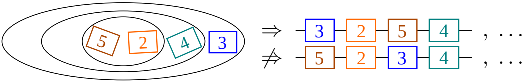


Consistency on Sequences

Models genome cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)	linear Euler	linear Euler	NP-comp. 3SAT3	NP-comp. com. int.	

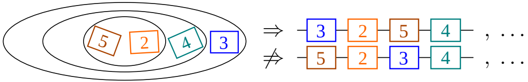
Consistency on Sequences

Nested Common Intervals on sequences



Consistency on Sequences

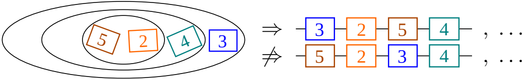
Nested Common Intervals on sequences



- NP-complete, even if
  - maximum cluster size = 5

Consistency on Sequences

Nested Common Intervals on sequences



Ján Manuch & Murray Patterson

- NP-complete, even if
    - maximum cluster size = 3
    - maximum multiplicity = 2
    - sum of multiplicities given

Consistency on Sequences

Models genome / cluster	unsigned adjacencies	signed adjacencies	common intervals	framed common intervals	nested common intervals
permutations (no dup's)	linear graph	linear graph	polyn. PQ trees	polyn. PQ trees	polyn. PQ trees
sequences (incl. dup's)	linear Euler	linear Euler	NP-comp. 3SAT3	NP-comp. com. int.	NP-comp. 3SAT3