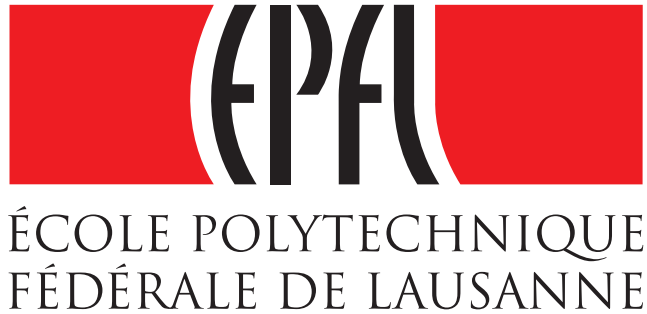


Principled Synteny Analysis

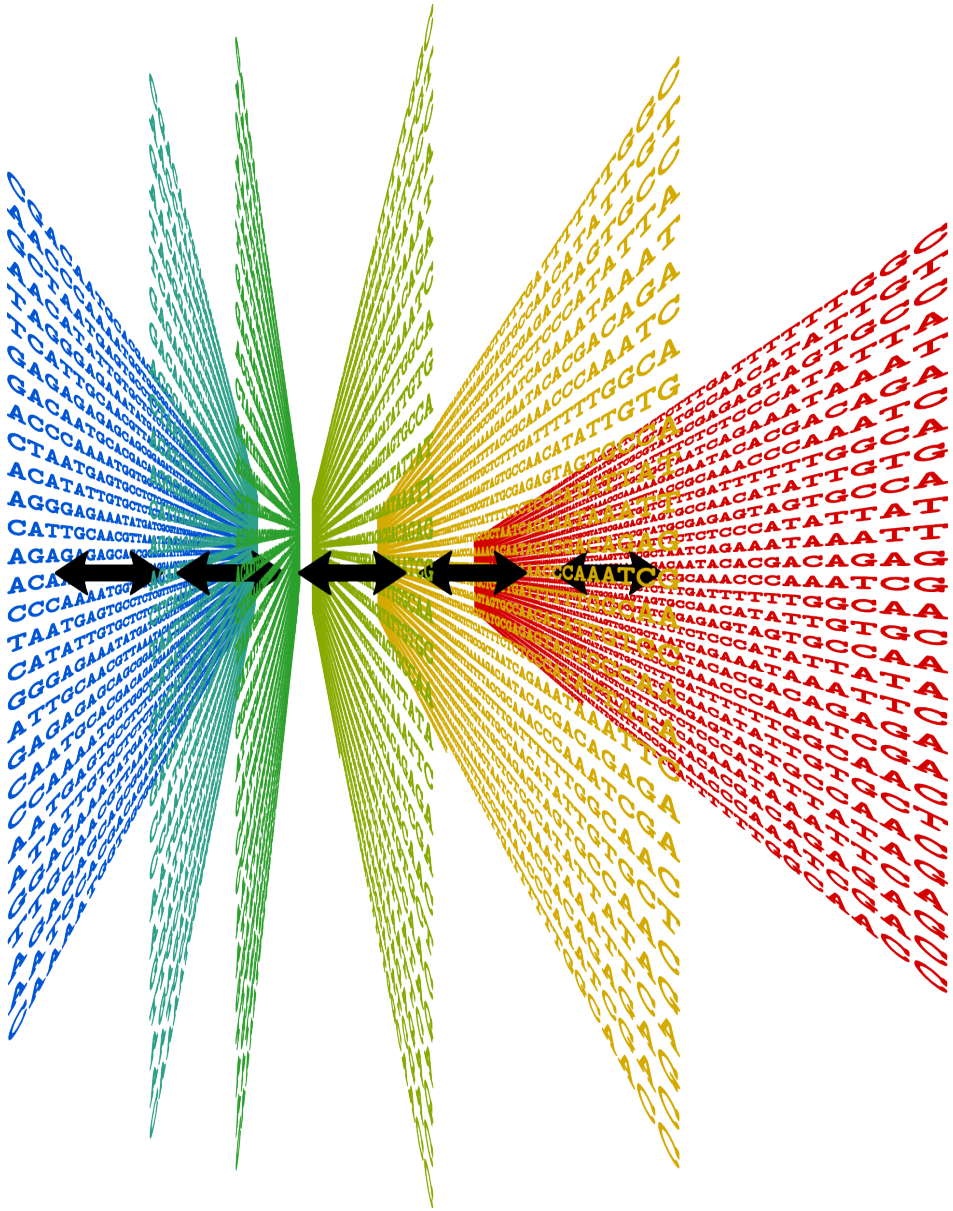
Daniel Doerr, Bernard M.E. Moret

School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

June 13, 2016

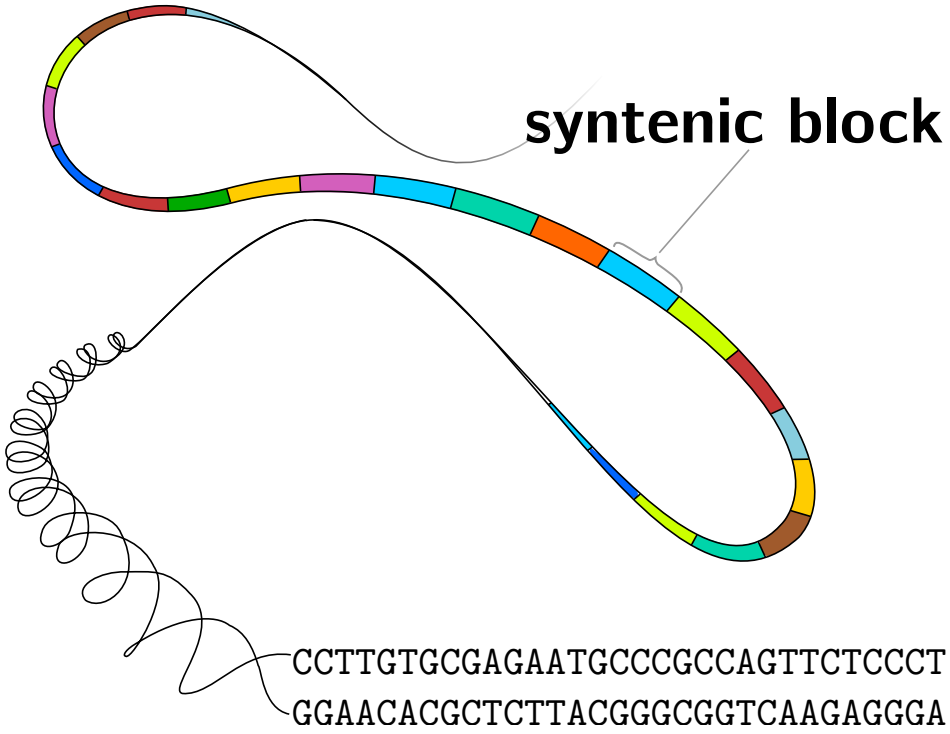
Comparing genomes

- Continued interest in large-scale whole-genome comparisons
- Base-by-base comparison infeasible for big genomes



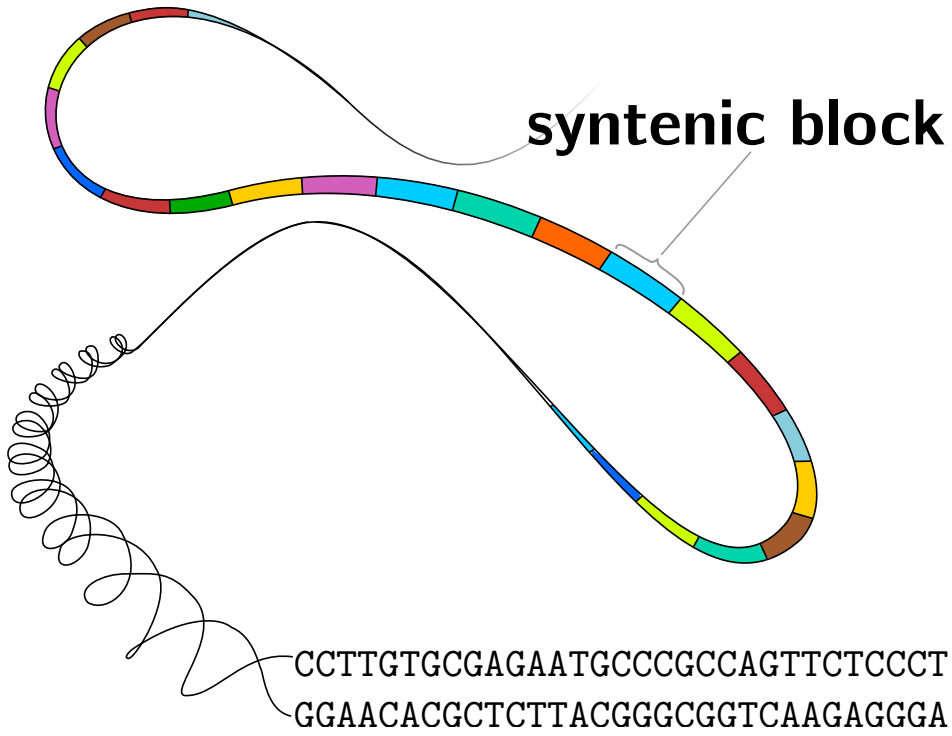
Abstraction by decomposition

- Genomes decomposed into *syntenic blocks*
- Essential for genome rearrangement studies



Abstraction by decomposition

- Genomes decomposed into *syntenic blocks*
- Essential for genome rearrangement studies



Protein-coding genes form basis for most rearrangement studies

$\Longrightarrow$ Omission of many other conserved genomic regions

What is synteny?

When is a syntenic block conserved? A zoo of definitions:

- *synteny*: Renwick, 1971: "*same ribbon*", co-location on same chromosome
- Collinear
- local rearrangements allowed
- Tool-centric: FISH, GRIMM/DRIMM-Synteny, Cyntenator, i-ADHoRe, Sibelia, CoGe, Satsuma, *etc.*

What is synteny?

When is a syntenic block conserved? A zoo of definitions:

- *synteny*: Renwick, 1971: "*same ribbon*", co-location on same chromosome
- Collinear
- local rearrangements allowed
- Tool-centric: FISH, GRIMM/DRIMM-Synteny, Cyntenator, i-ADHoRe, Sibelia, CoGe, Satsuma, *etc.*

Does there exist "one true syntenic block"?

$\Longrightarrow$ Different levels of granularity

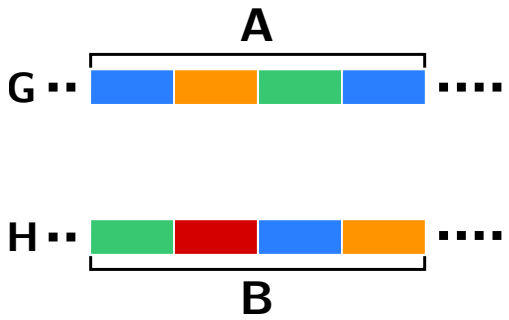
Synteny criterion

syntenic block (SB)

Set of contiguous syntenic blocks

homology assignment between blocks

Set $\mathcal{H}$ of pairwise equivalence relations



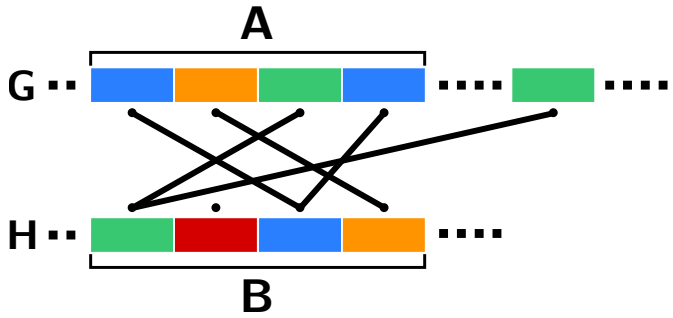
Synteny criterion

syntenic block (SB)

Set of contiguous syntenic blocks

homology assignment between blocks

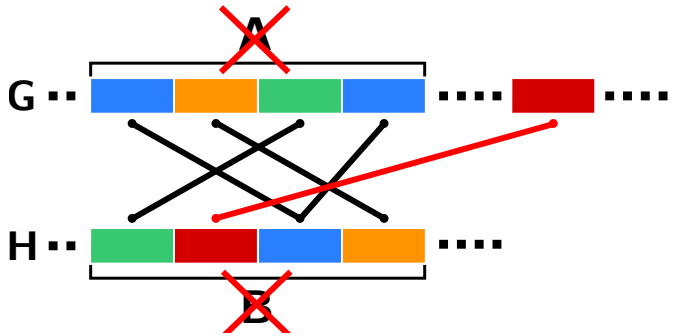
Set $\mathcal{H}$ of pairwise equivalence relations



homologous SB pair [Ghiurcuta and Moret, 2014]

Given two genomes G, H and homology assignment $\mathcal{H}$, two SBs $A \subseteq G$ and $B \subseteq H$ are homologous if for each

- $a \in A$:
  $\exists\ (a,h) \in \mathcal{H}, h \in H \implies (a,b') \in \mathcal{H}, b' \in B$
- $b \in B$:
  $\exists\ (b,g) \in \mathcal{H}, g \in G \implies (a',b) \in \mathcal{H}, a' \in A$



Synteny workflow

A workflow for principled synteny analysis

Genome segmentation

Apply synteny criterion to genome segmentation!

Genome segmentation

Apply synteny criterion to genome segmentation!

Sequence segmentation problem [Visnovská et al., 2013]

Set $\mathcal{S}$ of genomic subsequences is a **segmentation** of genomes $G_1, \ldots, G_k$ if for any $S, S' \in \mathcal{S}$ holds:

- for each site $X \in S$ s.t. $\nexists\, Y \in S'$ and X, Y homologous $\implies \nexists\, S'' \in \mathcal{S} \setminus \{S, S'\}$, s.t. $Z \in S''$ and X, Z homologous; id. for each site $X' \in S'$

```
S   :CG_GATG       AA
S'  :C_AGTTCATGATAAT
```

Genome segmentation

Apply synteny criterion to genome segmentation!

Sequence segmentation problem [Visnovská et al., 2013]

Set $\mathcal{S}$ of genomic subsequences is a **segmentation** of genomes $G_1, \ldots, G_k$ if for any $S, S' \in \mathcal{S}$ holds:

- for each site $X \in S$ s.t. $\nexists\, Y \in S'$ and X, Y homologous $\implies \nexists\, S'' \in \mathcal{S} \setminus \{S, S'\}$, s.t. $Z \in S''$ and X, Z homologous; id. for each site $X' \in S'$

```
S   :CG_GATG      AA
S'  :C_AGTTCATGATAAT
S'' :       ___TTA  TCATT
S''':       TT_ATT       TGCAA
```

Segmentation invalid!

Genome segmentation

Apply synteny criterion to genome segmentation!

Sequence segmentation problem [Visnovská et al., 2013]

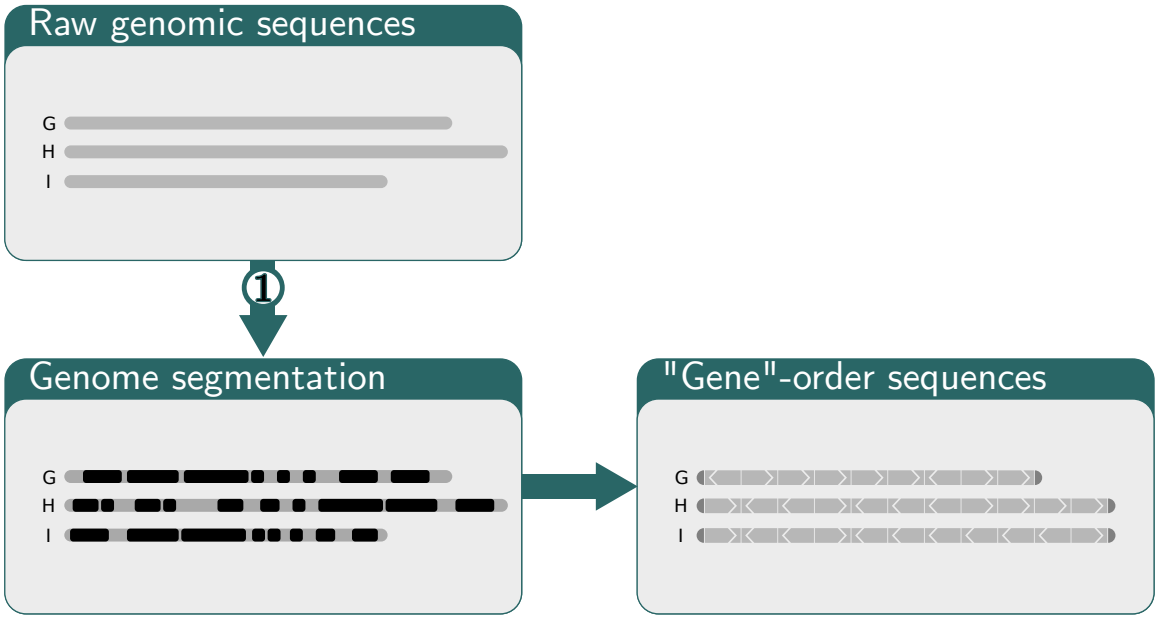
Set $\mathcal{S}$ of genomic subsequences is a **segmentation** of genomes $G_1, \ldots, G_k$ if for any $S, S' \in \mathcal{S}$ holds:

- for each site $X \in S$ s.t. $\nexists\, Y \in S'$ and X, Y homologous $\implies \nexists\, S'' \in \mathcal{S} \setminus \{S, S'\}$, s.t. $Z \in S''$ and X, Z homologous; id. for each site $X' \in S'$
- non-overlapping, if S, S' belong to same genome

```
S   :CG_GATG      AA
S'  :C_AGTTCATGATAAT
S'' :       ___TTA  TCATT
S''':       TT_ATT       TGCAA
```

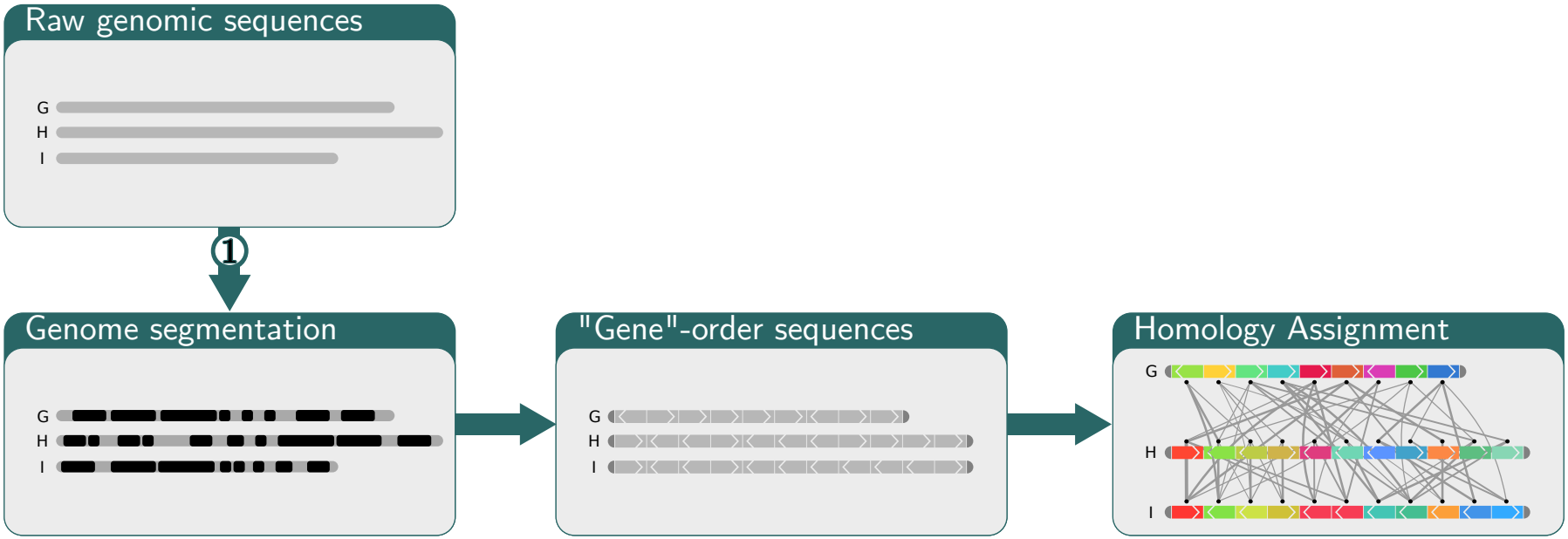
Segmentation invalid!

Synteny workflow



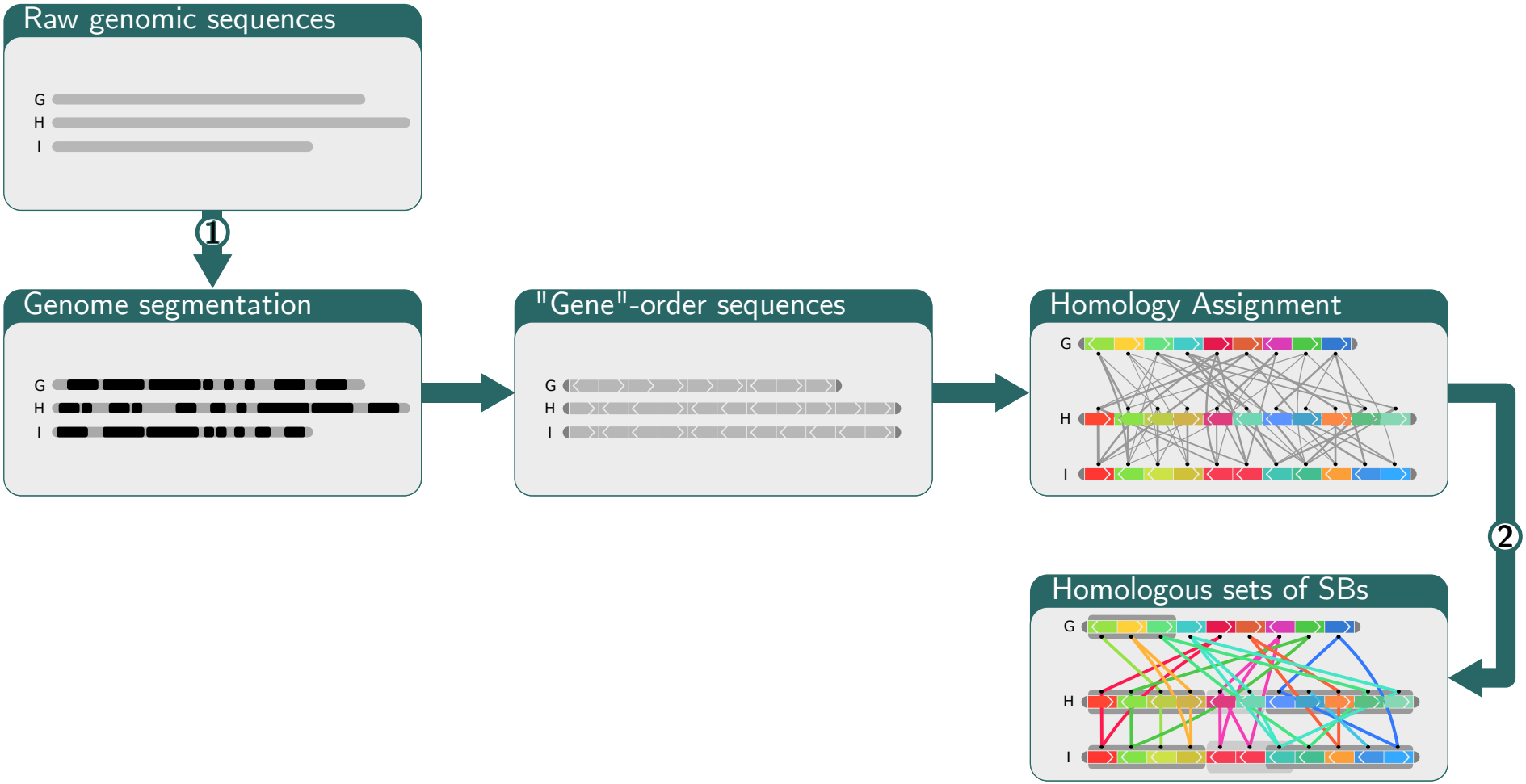
Synteny workflow

LCBB

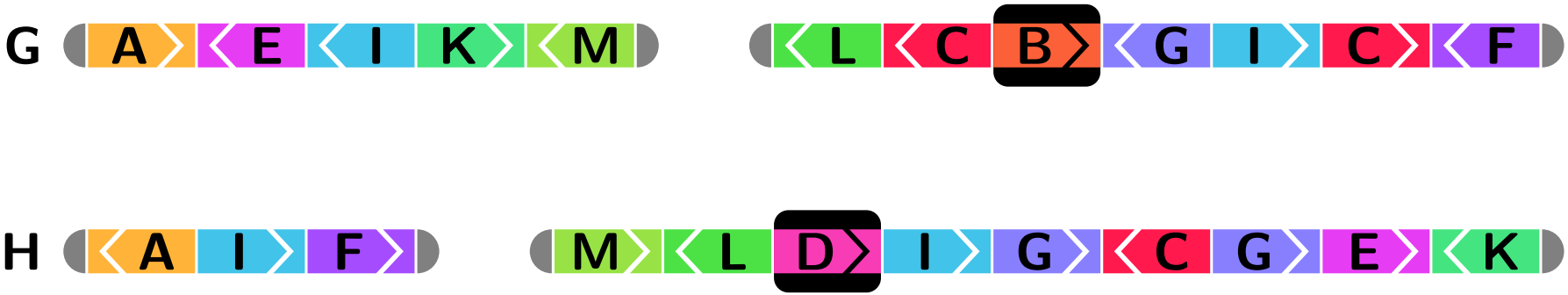


Synteny workflow

LCBB



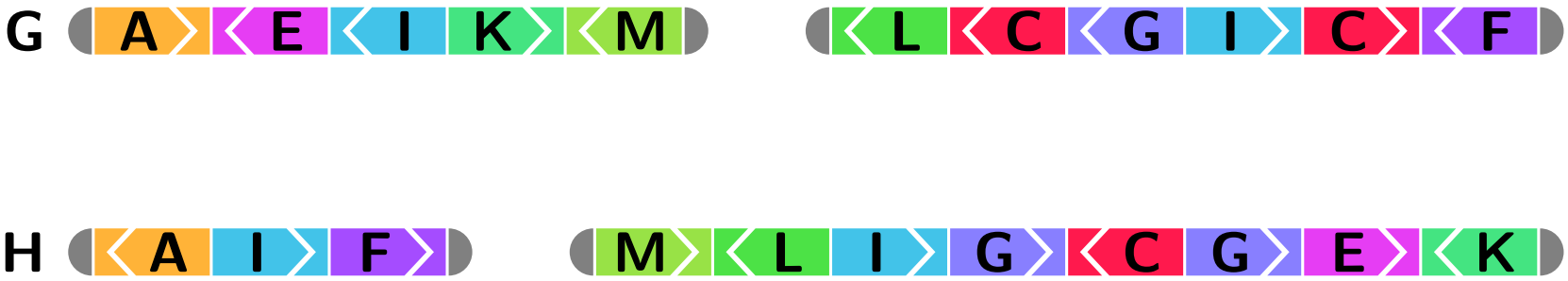
Discovery of homologous SBs



Remove all markers not shared between **G** and **H**

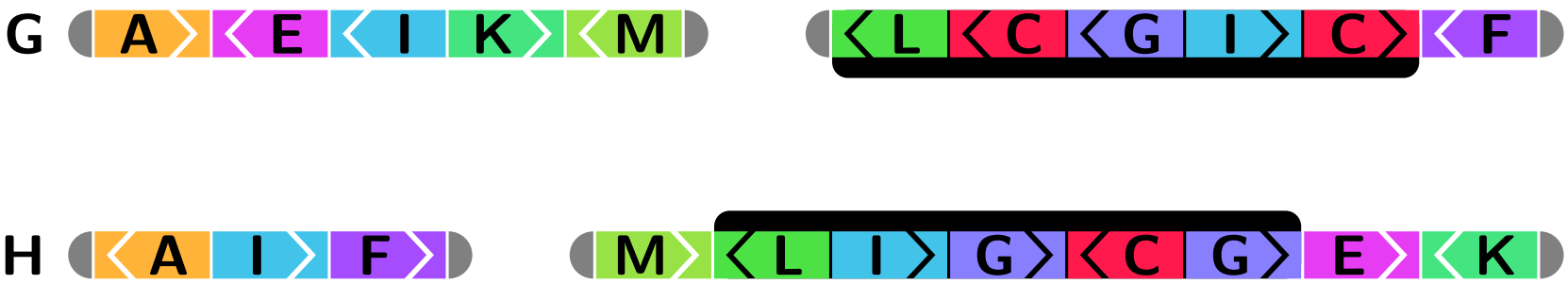
Discovery of homologous SBs

LCBB



Discovery of homologous SBs

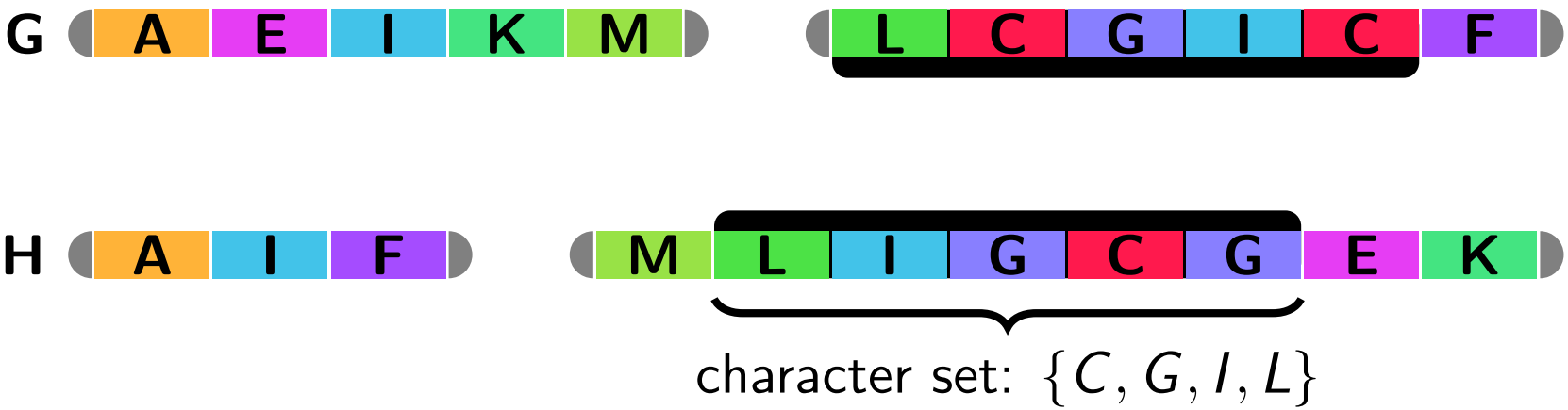
LCBB



Two blocks are syntenic if each marker in both blocks is homologous to at least one marker of the other block.

Discovery of homologous SBs

LCBB



common intervals: intervals with identical character set

[Uno and Yagiura, 2000, Schmidt and Stoye, 2004]

Non-transitive homology assignments

LCBB

- clustering of genomic markers into homologous groups is error-prone
- Synteny can is a powerful filter for false positive homologies

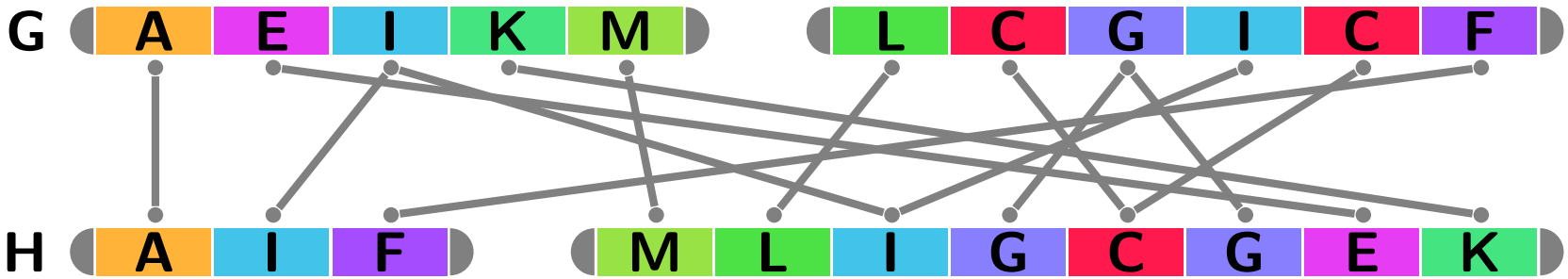
Non-transitive homology assignments

- clustering of genomic markers into homologous groups is error-prone
- Synteny can is a powerful filter for false positive homologies
- Omit clustering by relaxing homology:

non-transitive homology assignment

Let Σ be the universe of markers. A *nontransitive homology assignment* is a *reflexive* and *symmetric* relation, $\tilde{\mathcal{H}} \subseteq \Sigma \times \Sigma$

Discovery of non-transitive homologous SBs

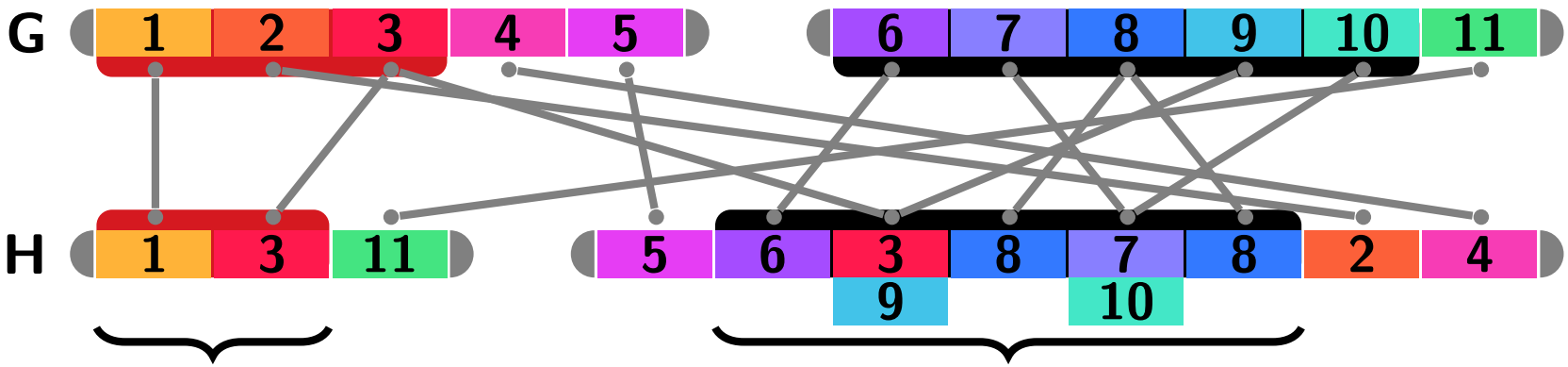


Homology statements: Links between characters

Discovery of non-transitive homologous SBs

LCBB

How many additional homology statements are needed to make an SB pair homologous?



1-approximate, common character set: $\{1, 3\}$

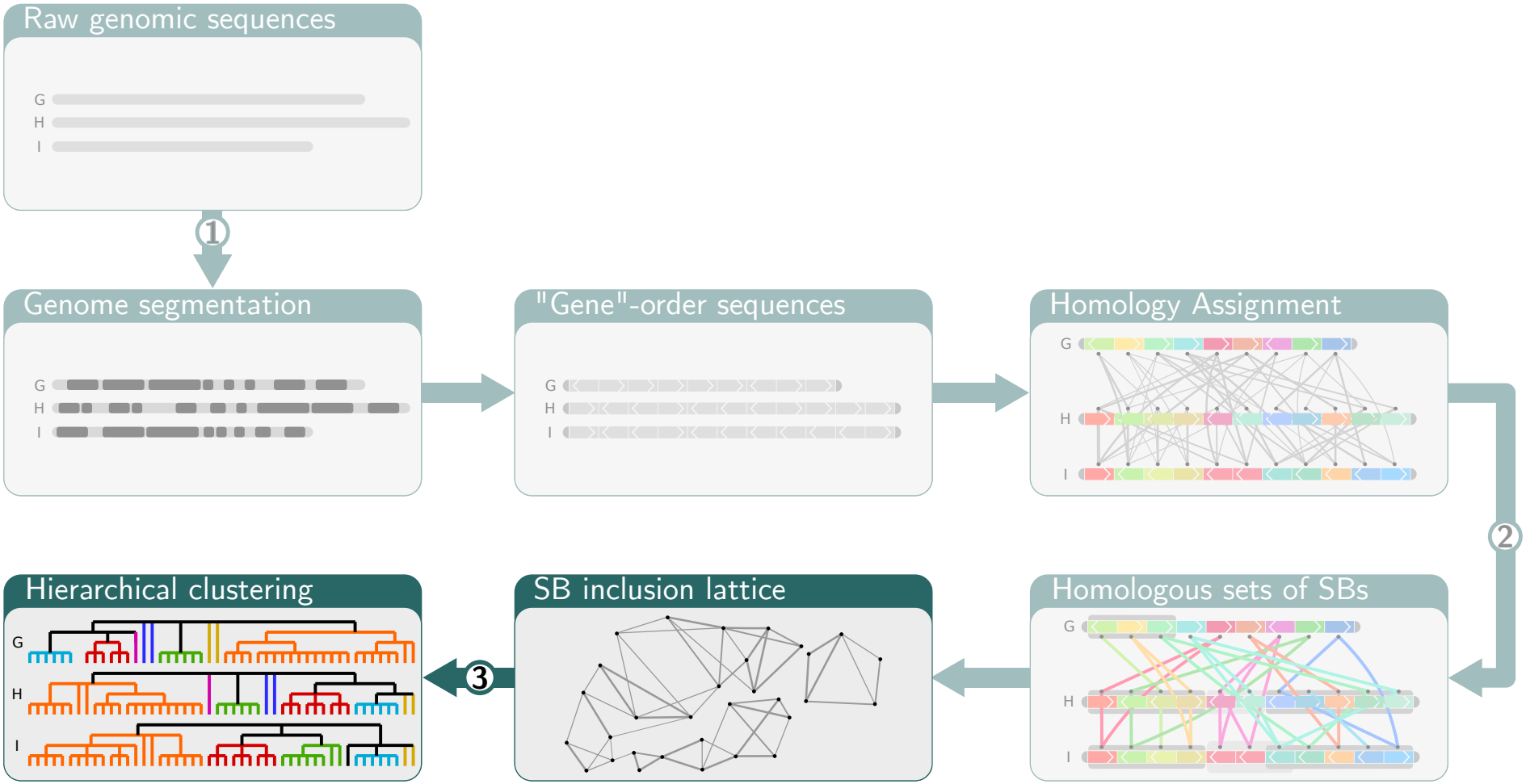
common character set: $\{6, 7, 8, 9, 10\}$

δ-approximate weak common intervals in indeterminate strings

[Doerr et al., 2014]

Synteny hierarchy construction

LCBB



SB inclusion lattice

LCBB

SB inclusion lattice:

- vertex: syntenic block
- directed **black** edge: $v_A \rightarrow v_B$ – syntenic block A subinterval of B
- undirected **colored** edge: v_A—v_B – A is homologous to B

SB hierarchy

LCBB

A set of intervals is **commuting** if for any two of its members A, B holds:

- A subinterval of B, or
- B subinterval of A, or
- A and B are disjoint

SB hierarchy

A set of intervals is **commuting** if for any two of its members A, B holds:

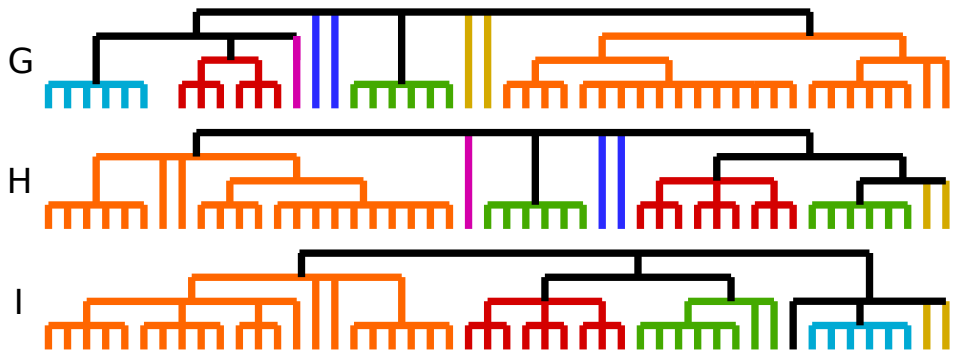
- A subinterval of B, or
- B subinterval of A, or
- A and B are disjoint

Determine commuting sets of SBs for each genome independently!

SB hierarchy

A set of intervals is **commuting** if for any two of its members A, B holds:

- A subinterval of B, or
- B subinterval of A, or
- A and B are disjoint



Determine commuting sets of SBs for each genome independently!

PriSy

PriSy* – a tool for Principled Synteny Analysis

*working title, release: end of this year

Thank you!

References

[Doerr et al., 2014] Doerr, D., Stoye, J., Böcker, S., and Jahn, K. (2014).
Identifying gene clusters by discovering common intervals in indeterminate strings.
BMC Genomics, 15(Suppl 6):S2.

[Ghiurcuta and Moret, 2014] Ghiurcuta, C. G. and Moret, B. M. E. (2014).
Evaluating synteny for improved comparative studies.
Bioinformatics, 30(12):i9–18.

[Schmidt and Stoye, 2004] Schmidt, T. and Stoye, J. (2004).
Quadratic time algorithms for finding common intervals in two and more sequences.
In *Proc. of CPM 2004*, volume 3109 of *LNCS*, pages 347–358.

[Uno and Yagiura, 2000] Uno, T. and Yagiura, M. (2000).
Fast algorithms to enumerate all common intervals of two permutations.
Algorithmica, 26(2):290–309.

[Visnovská et al., 2013] Visnovská, M., Vinař, T., and Brejová, B. (2013).
DNA Sequence Segmentation Based on Local Similarity.
In *Proc. of ITAT*, volume 1003, pages 36–43.