More ideas about new orthology inference methods

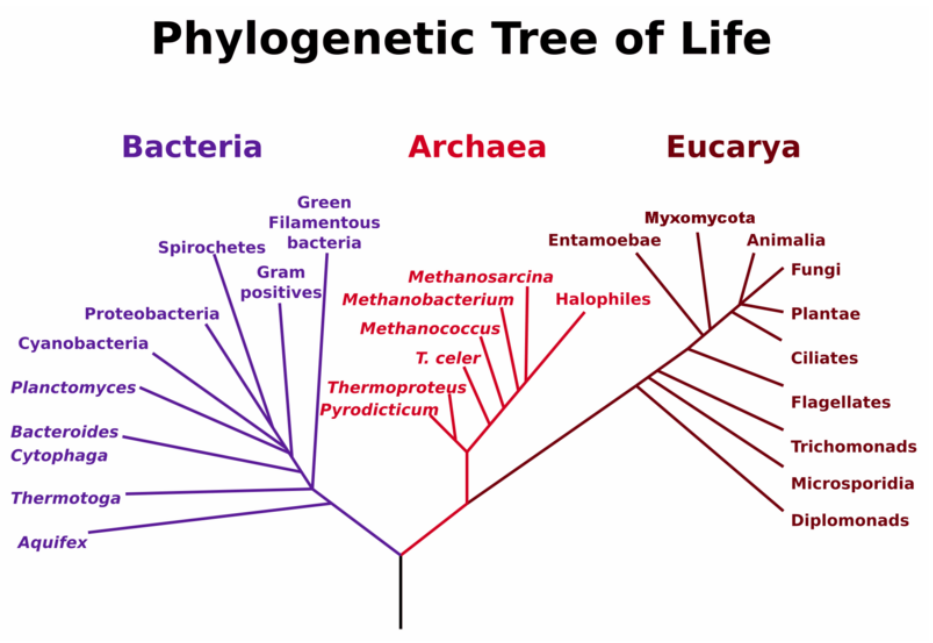
Manuela Geiß

Bioinformatics Group
University of Leipzig

34th TBI Winterseminar
Bled, 12th February 2019

Why Orthology Analysis?

Orthology analysis is an important part of data analysis in many areas such as comparative genomics and molecular phylogenetics



source: https://en.wikipedia.org/wiki/Tree_of_life_(biology)

Idea: There is only one **true** tree of life – we just need good methods to detect it!

Tree-based vs. graph-based methods

Tree-based:

- species tree must be known, gene tree via sequence alignments
  $\rightarrow$ tree reconciliation gives orthology relation
- accuracy highly depends on quality of trees
- high computational costs

Graph-based:

- construction of the trees from sequence data
- lower computational costs
- many tools restricted to small number of species (except `ProteinOrtho`[1])
- some tools even include manual correction

[1] Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ, 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124

Tree-based vs. graph-based methods

Tree-based:

- species tree must be known, gene tree via sequence alignments
  $\rightarrow$ tree reconciliation gives orthology relation
- accuracy highly depends on quality of trees
- high computational costs

Graph-based:

- construction of the trees from sequence data
- lower computational costs
- many tools restricted to small number of species (except `ProteinOrtho`[1])
- some tools even include manual correction

$\rightarrow$ Our overall-goal: improve orthology inference/develop new methods

[1] Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ, 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124

What are best matches?

True divergence times of genes/species often not known $\rightarrow$ many tools use Best Match Heuristics

What are best matches?

True divergence times of genes/species often not known $\rightarrow$ many tools use Best Match Heuristics

Definition

The sequence y of species Y is a best match of the sequence x of species X if y is "closest" to x among all genes in species Y.

What are best matches?

True divergence times of genes/species often not known $\rightarrow$ many tools use Best Match Heuristics

Definition

The sequence y of species Y is a best match of the sequence x of species X if y is "closest" to x among all genes in species Y.

Definition

The sequences x and y are reciprocal best matches if y is closest to x and x is closest to y.

What are best matches?

True divergence times of genes/species often not known → many tools use Best Match Heuristics

Definition

The sequence y of species Y is a best match of the sequence x of species X if y is "closest" to x among all genes in species Y.

Definition

The sequences x and y are reciprocal best matches if y is closest to x and x is closest to y.

→ Goal: Deeper understanding of (reciprocal) Best Match Graphs to make the process more efficient

Best Match Graphs (BMGs)

here: "closest" = closest last common ancestor (lca)

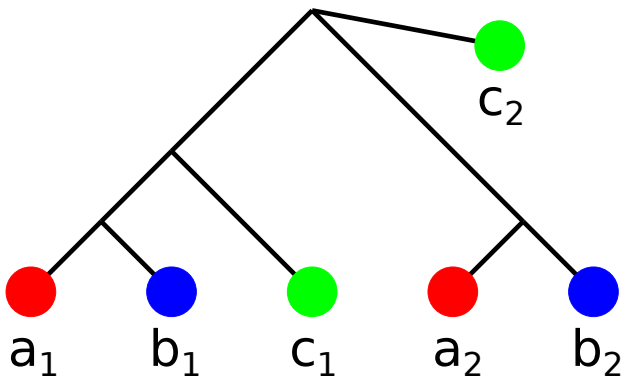
Definition

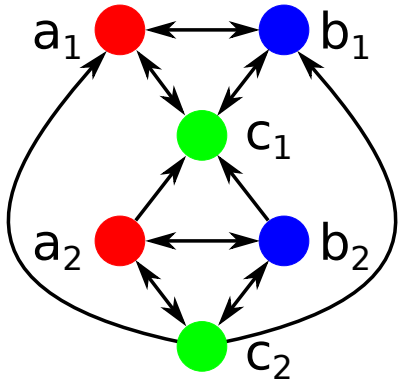
The leaf y is a best match of the leaf x in a tree T if $\sigma(x) \neq \sigma(y)$, and

(i) $\mathrm{lca}(x, y) \preceq \mathrm{lca}(x, y')$ for all leaves y' from species $\sigma(y') = \sigma(y)$.

We write $x \to y$.

σ = colors (= species)





Reciprocal Best Match Graphs (RBMGs)

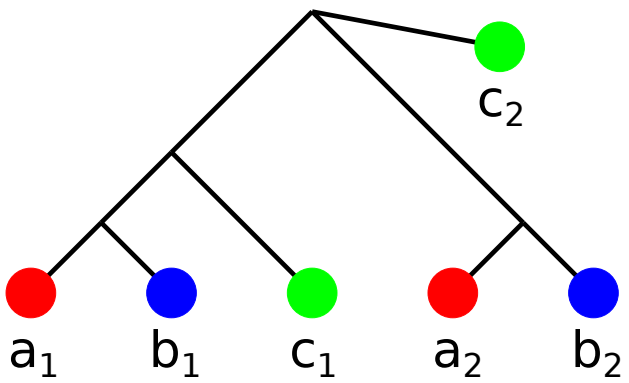
Definition

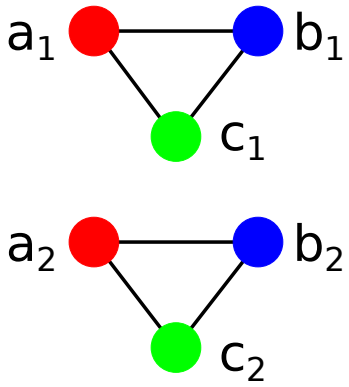
The leaf y is a reciprocal best match of the leaf x in a tree T if $\sigma(x) \neq \sigma(y)$, and

(i) $\mathrm{lca}(x, y) \preceq \mathrm{lca}(x, y')$ for all leaves y' from species $\sigma(y') = \sigma(y)$, and

(ii) $\mathrm{lca}(x, y) \preceq \mathrm{lca}(y, x')$ for all leaves x' from species $\sigma(x') = \sigma(x)$.

σ = colors (= species)





Reciprocal Best Match Graphs (RBMGs)

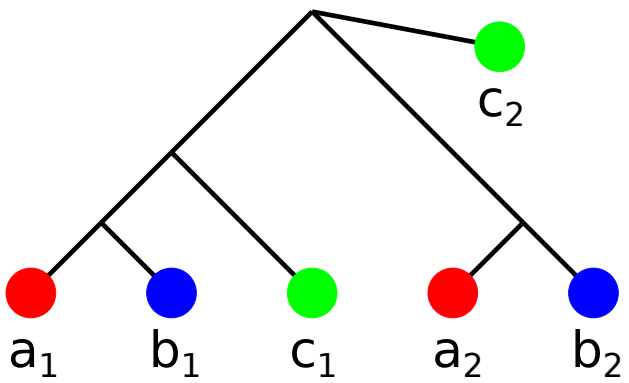
Definition

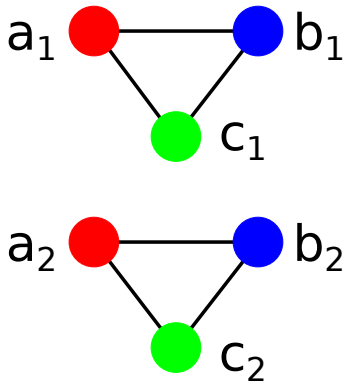
The leaf y is a reciprocal best match of the leaf x in a tree T if $\sigma(x) \neq \sigma(y)$, and

(i) $\mathrm{lca}(x, y) \preceq \mathrm{lca}(x, y')$ for all leaves y' from species $\sigma(y') = \sigma(y)$, and

(ii) $\mathrm{lca}(x, y) \preceq \mathrm{lca}(y, x')$ for all leaves x' from species $\sigma(x') = \sigma(x)$.

σ = colors (= species)





→ Which (un-)directed graphs are (Reciprocal) Best Match Graphs, i.e., have a tree representation?

Mathematical Results about (Reciprocal) Best Match Graphs

BMGs[1]

- Two characterizations for 2-cBMGs via triples and neighborhoods
  $\rightarrow$ Recognition in polynomial time
- Characterization for n-cBMGs via Aho-Tree from 2-cBMGs
  $\rightarrow$ Recognition and tree reconstruction in polynomial time
- Unique least resolved tree

[1] M. Geiß, E. Chavez, M. Gonzalez, A. Lopez, D. Valdivia, M. H. Rosales, B.M.R. Stadler, Marc Hellmuth, P.F. Stadler, 2018, Best Match Graphs, J. Math. Biology *(accepted - to appear)*

[2] M. Geiß, Marc Hellmuth, P.F. Stadler, 2019, Reciprocal Best Match Graphs.. *(manuscript in preparation)*

Mathematical Results about (Reciprocal) Best Match Graphs

BMGs[1]

- Two characterizations for 2-cBMGs via triples and neighborhoods
  $\rightarrow$ Recognition in polynomial time
- Characterization for n-cBMGs via Aho-Tree from 2-cBMGs
  $\rightarrow$ Recognition and tree reconstruction in polynomial time
- Unique least resolved tree

RBMGs[2]

- Classification of three distinct groups of 3-cRBMGs
  $\rightarrow$ Recognition in polynomial time
- Characterization for n-cRBMGs via supertree from 3-cRBMGs
  $\rightarrow$ Recognition and tree reconstruction presumably **not** in polynomial time
- No unique least resolved tree

[1] M. Geiß, E. Chavez, M. Gonzalez, A. Lopez, D. Valdivia, M. H. Rosales, B.M.R. Stadler, Marc Hellmuth, P.F. Stadler, 2018, Best Match Graphs, J. Math. Biology *(accepted - to appear)*

[2] M. Geiß, Marc Hellmuth, P.F. Stadler, 2019, Reciprocal Best Match Graphs.. *(manuscript in preparation)*

Mathematical Results about (Reciprocal) Best Match Graphs

BMGs[1]

- Two characterizations for 2-cBMGs via triples and neighborhoods
  $\rightarrow$ Recognition in polynomial time
- Characterization for n-cBMGs via Aho-Tree from 2-cBMGs
  $\rightarrow$ Recognition and tree reconstruction in polynomial time
- Unique least resolved tree

RBMGs[2]

- Classification of three distinct groups of 3-cRBMGs
  $\rightarrow$ Recognition in polynomial time
- Characterization for n-cRBMGs via supertree from 3-cRBMGs
  $\rightarrow$ Recognition and tree reconstruction presumably **not** in polynomial time
- No unique least resolved tree
  $\rightarrow$ Much information lost by only looking at RBMGs!

[1] M. Geiß, E. Chavez, M. Gonzalez, A. Lopez, D. Valdivia, M. H. Rosales, B.M.R. Stadler, Marc Hellmuth, P.F. Stadler, 2018, Best Match Graphs, J. Math. Biology *(accepted - to appear)*

[2] M. Geiß, Marc Hellmuth, P.F. Stadler, 2019, Reciprocal Best Match Graphs.. *(manuscript in preparation)*

How can we use all this?

Theorem [1]

In pure DL scenarios (i.e. in the absence of HGT events) the reciprocal best match graph can only produce false positive but not false negative orthology assignments.
$\Rightarrow$ *The true orthology relation has to be contained in the RBMG.*

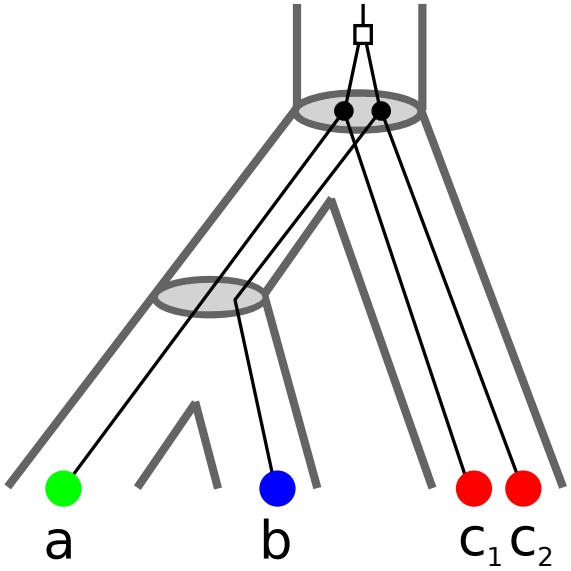
[1] M. Geiß, A. Lopez, D. Valdivia, M. H. Rosales, Marc Hellmuth, P.F. Stadler, 2019, Best Match Graphs and Reconciliation of Gene Trees with Species Trees. J. Math. Biology *(manuscript in preparation)*

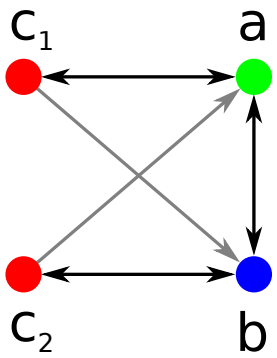
How can we use all this?

Theorem [1]

In pure DL scenarios (i.e. in the absence of HGT events) the reciprocal best match graph can only produce false positive but not false negative orthology assignments.
$\Rightarrow$ *The true orthology relation has to be contained in the RBMG.*

$\rightarrow$ **Some** false positive edges can be identified using best match graphs

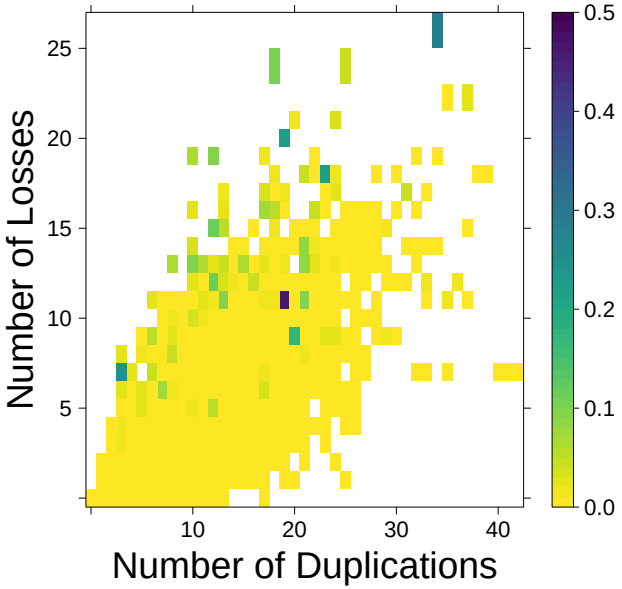




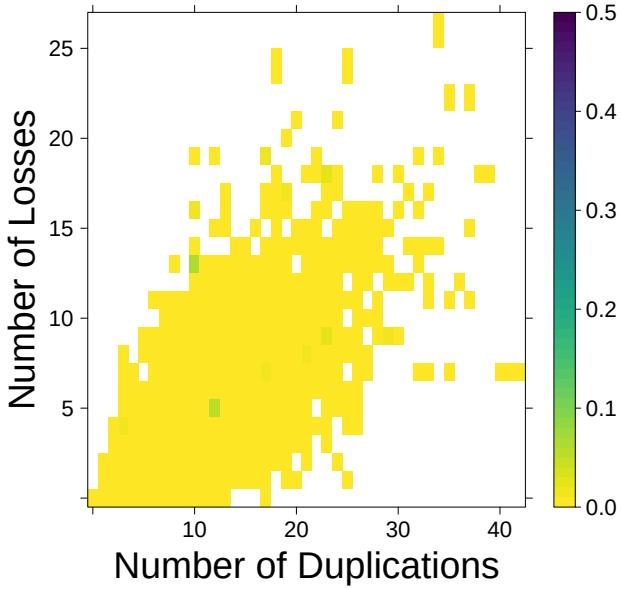
Remove middle edge:
P_4-Editing (P4E)

[1] M. Geiß, A. Lopez, D. Valdivia, M. H. Rosales, Marc Hellmuth, P.F. Stadler, 2019, Best Match Graphs and Reconciliation of Gene Trees with Species Trees. J. Math. Biology *(manuscript in preparation)*

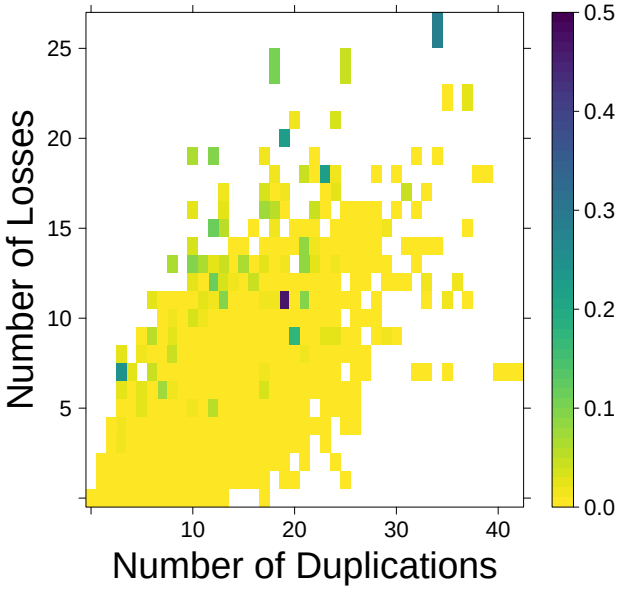
Simulation results with 0 HGT events



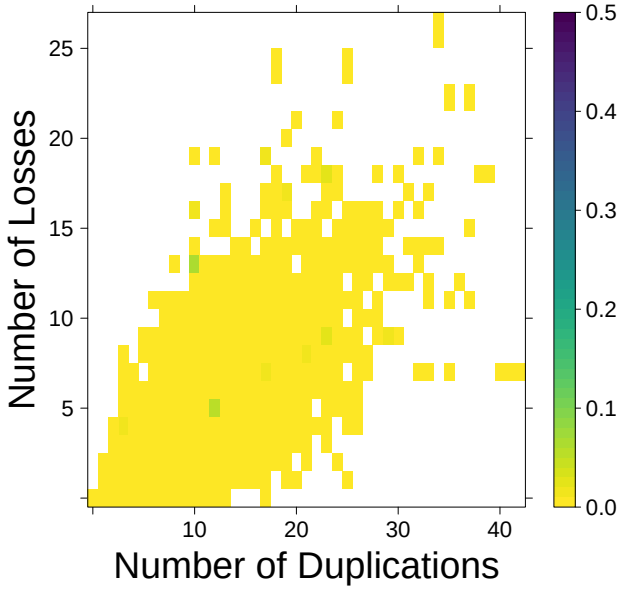
FPR
before
vs.
after
P4E

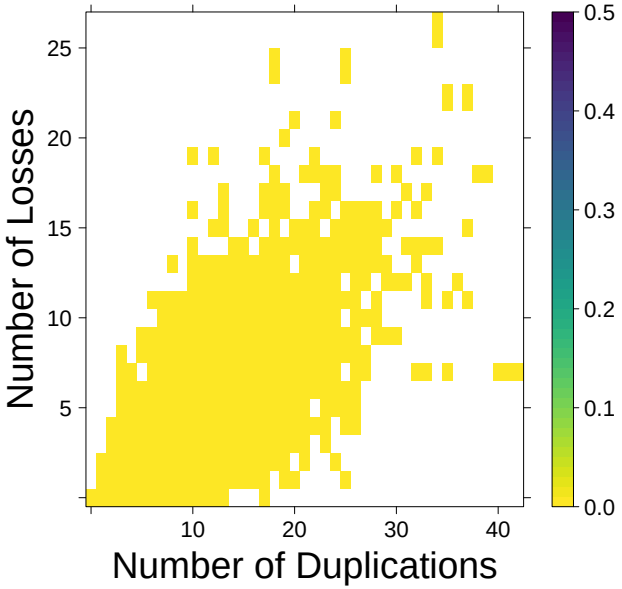


Simulation results with 0 HGT events

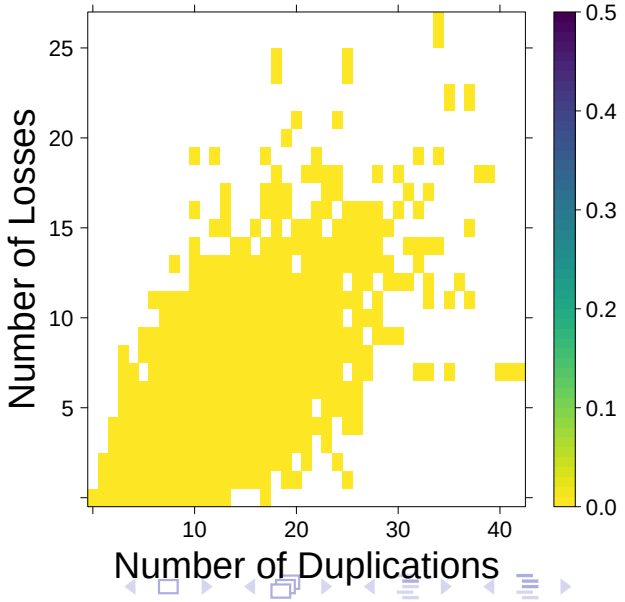


FPR
before
vs.
after
P4E

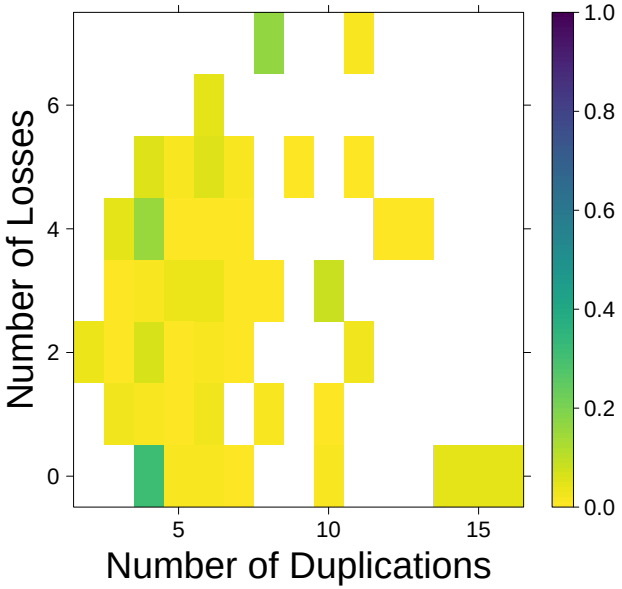




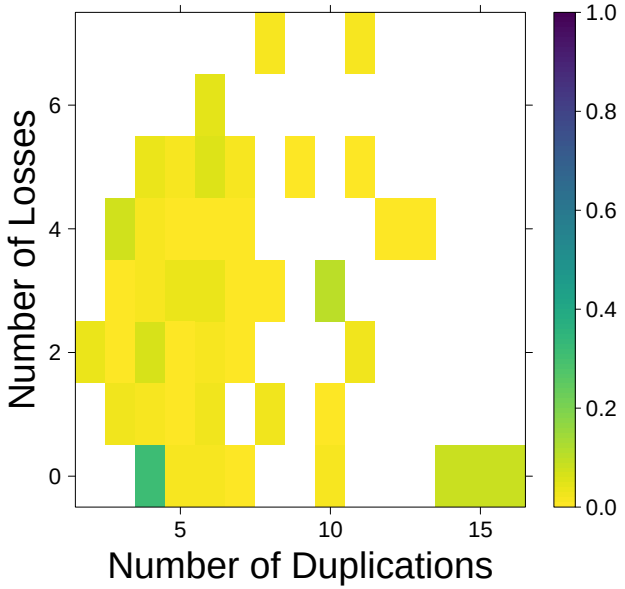
FNR
before
vs.
after
P4E



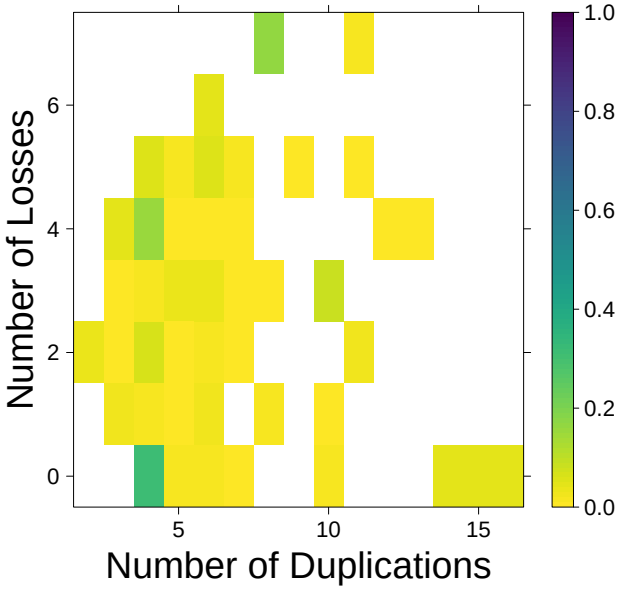
Simulation results with 1 HGT event



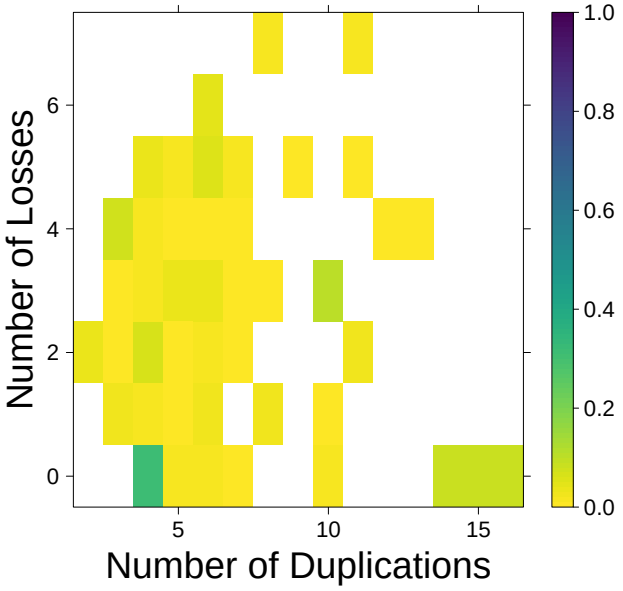
FPR
before
vs.
after
P4E

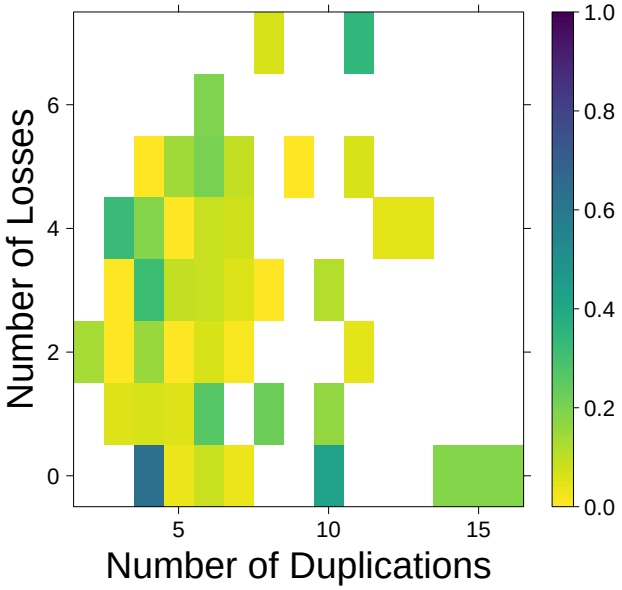


Simulation results with 1 HGT event

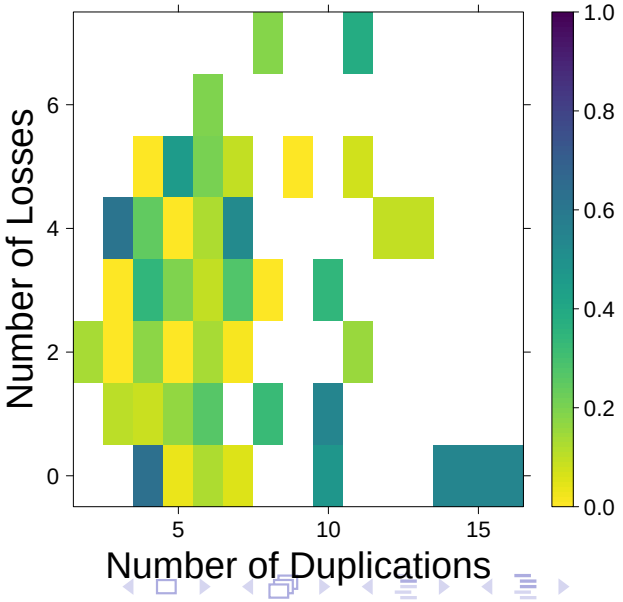


FPR
before
vs.
after
P4E

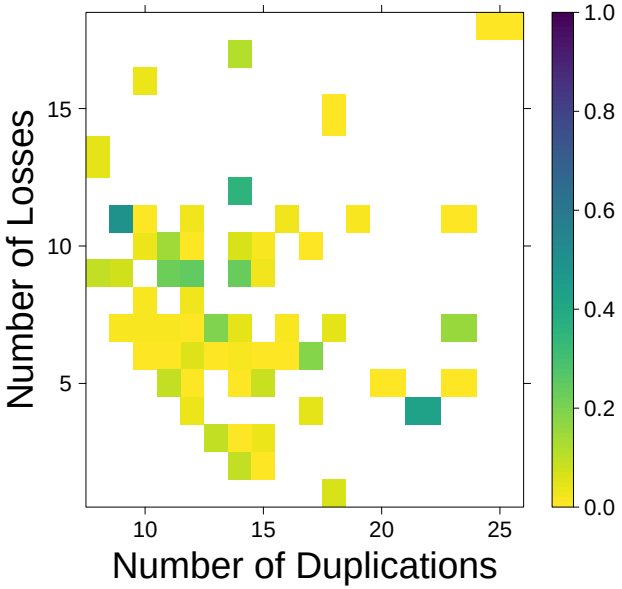




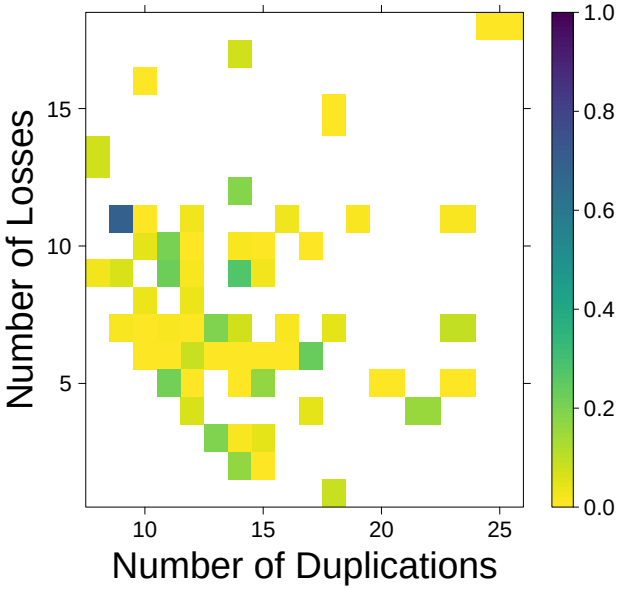
FNR
before
vs.
after
P4E

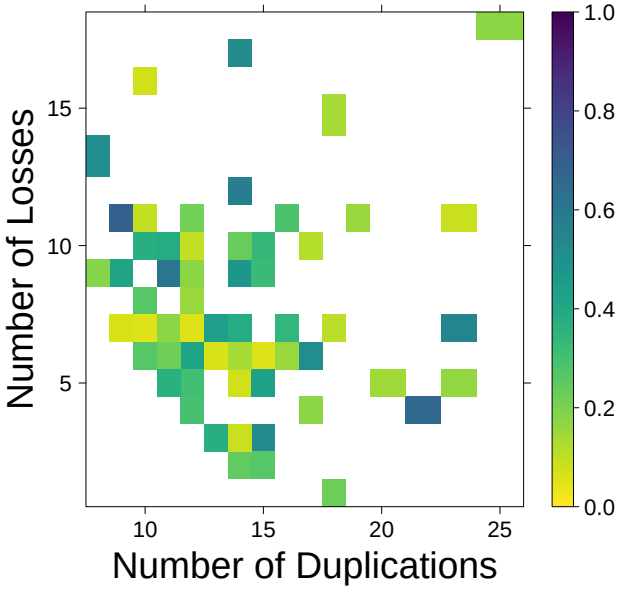


Simulation results with 4 HGT events

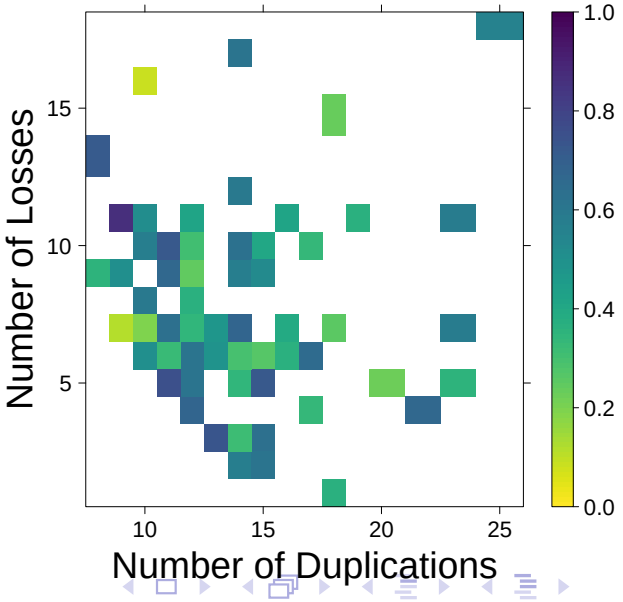


FPR
before
vs.
after
P4E





FNR
before
vs.
after
P4E



Summary & Outlook

Results so far:

- Characterization and tree reconstruction algorithms for BMGs and RBMGs
- RBMG contains no false positive orthologs in the absence of HGT
- P_4-Editing in the absence of HGT
- RBMG contains false negative orthologs in the presence of HGT

$\rightarrow$ RBMG loses much information that is still contained in the BMG!

Summary & Outlook

Results so far:

- Characterization and tree reconstruction algorithms for BMGs and RBMGs
- RBMG contains no false positive orthologs in the absence of HGT
- P_4-Editing in the absence of HGT
- RBMG contains false negative orthologs in the presence of HGT

$\rightarrow$ RBMG loses much information that is still contained in the BMG!

Next steps:

- BMGs might help to detect HGT events
- Improved graph editing based on characterization of BMGs and RBMGs

Special Thanks to:
Peter F. Stadler
Marc Hellmuth
Nicolas Wieseke
Edgar Chávez
Marcos González
Maribel Hernández Rosales
Alitzel López
Dulce Valdivia

Thank you
for your attention!

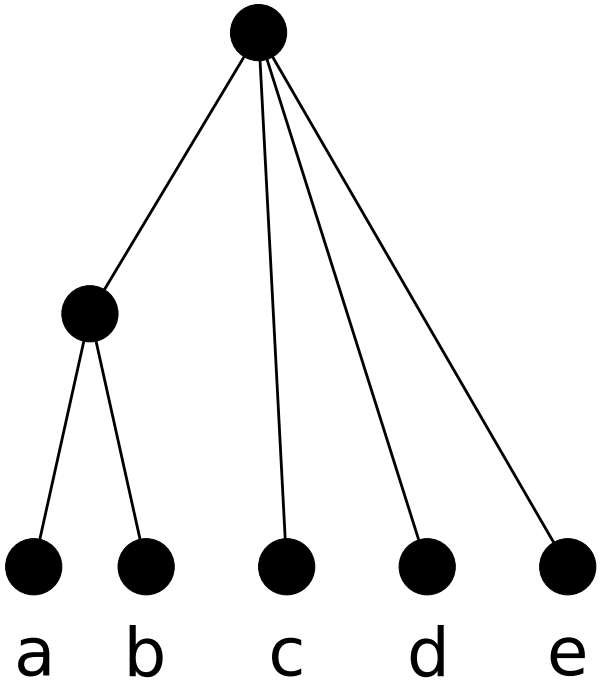
UNIVERSITÄT LEIPZIG

de NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

DAAD

Some basics: Rooted Trees and Triples

Rooted Tree T:



acyclic, connected graph

Triples:

- T *displays* a triple $ab|c$ if the path from c to the root is not intersected by the path from a to b.
- $\mathcal{R}(T) = \{ab|c, ab|d, ab|e\}$
- A set of triples R is said to be *consistent* if there is a tree T with $R \subseteq \mathcal{R}(T)$.
- Consistency-check via `BUILD`-algorithm in polynomial time. In case of consistency, it returns a tree T (the "Aho Tree") with $R \subseteq \mathcal{R}(T)$.

How do n-cBMGs look like?

Theorem

A colored digraph (G, σ) is a n-cBMG if and only if all induced subgraphs on two colors are 2-cBMG's and the union of the triples obtained from their least resolved trees forms a consistent set.

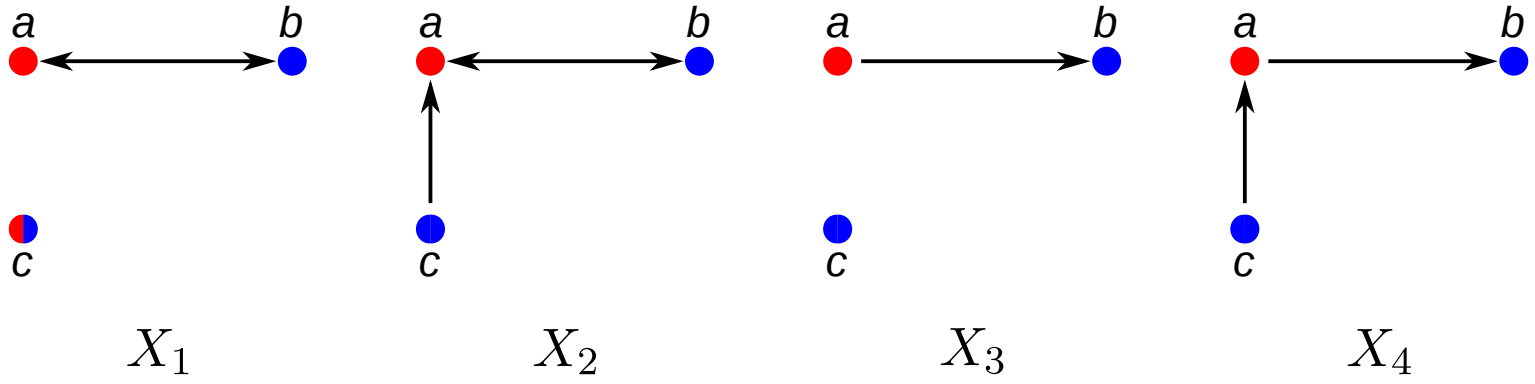
least resolved = "lowest possible resolution"

$\rightarrow$ The unique least resolved tree for (G, σ) can be reconstructed in cubic time

$\rightarrow$ All information that is needed, is contained in the 2-colored best match graphs!

The case of two colors: Characterization via triples

Some 2-colored subgraphs on 3 vertices give us constraints on the tree topology:



X_1, X_2, X_3, and X_4 all give the informative triple $ab|c$.

Theorem

A connected 2-colored digraph (G, σ) is a 2-cBMG if and only if $(G, \sigma) = G(Aho(\mathcal{R}(G, \sigma)), \sigma)$, where $\mathcal{R}(G, \sigma)$ is the set of all informative triples of (G, σ).

The case of two colors: Characterization via out-neighborhoods

Augenkrätze-Theorem

A connected 2-colored digraph (G, σ) is a 2-cBMG if and only if (G, σ) satisfies properties (N0), (N1), (N2), *and* (N3). *Moreover, the tree T defined by the $\mathcal{H}' := \{R'(\alpha) \mid \alpha \in \mathcal{N}\}$ is the unique least resolved tree that explains (G, σ).*

(N0) $\beta \subseteq N(\alpha)$ or $\beta \cap N(\alpha) = \emptyset$

(N1) $\alpha \cap N(\beta) = \beta \cap N(\alpha) = \emptyset$ implies
$N(\alpha) \cap N(N(\beta)) = N(\beta) \cap N(N(\alpha)) = \emptyset$.

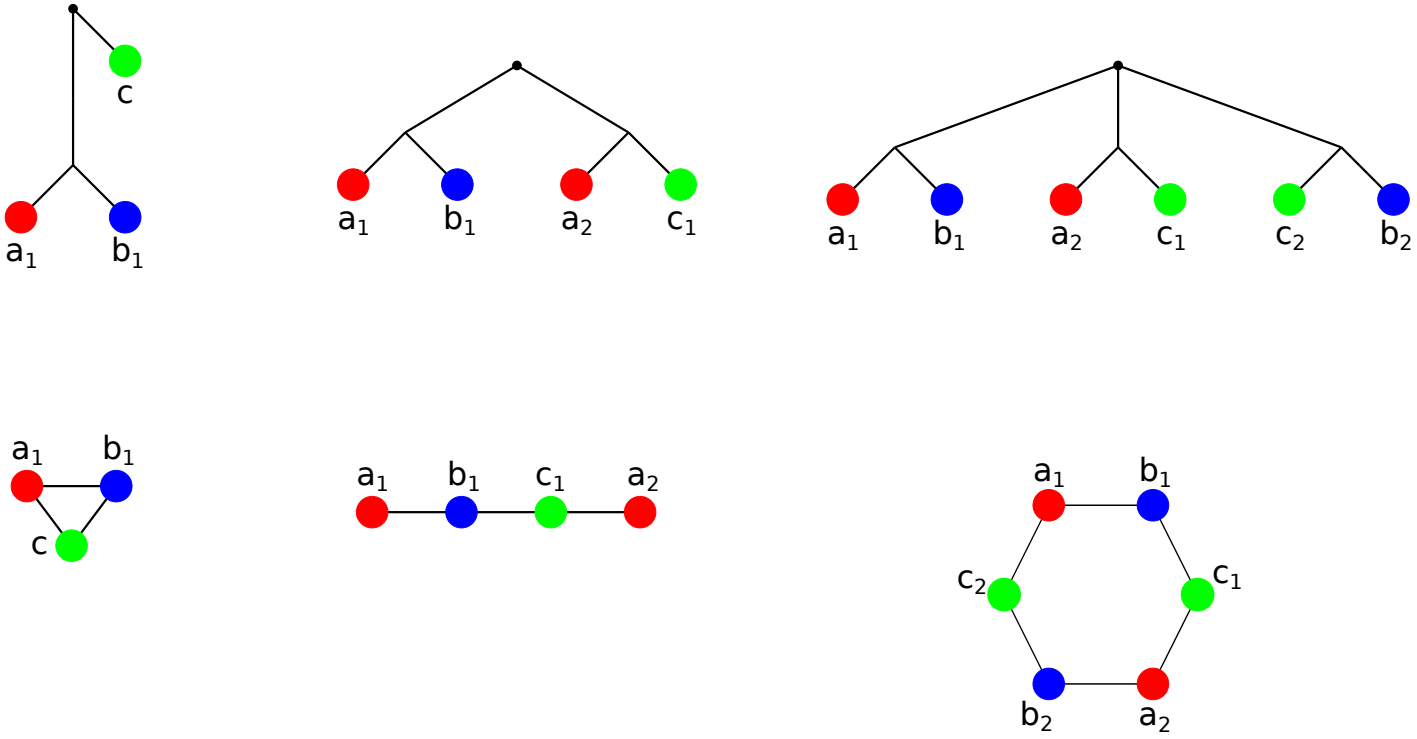
(N2) $N(N(N(\alpha))) \subseteq N(\alpha)$

(N3) $\alpha \cap N(N(\beta)) = \beta \cap N(N(\alpha)) = \emptyset$ and $N(\alpha) \cap N(\beta) \neq \emptyset$
implies $N^-(\alpha) = N^-(\beta)$ and $N(\alpha) \subseteq N(\beta)$ or $N(\beta) \subseteq N(\alpha)$

properties can be nicely checked by an algorithm

The three classes of 3-cRBMGs

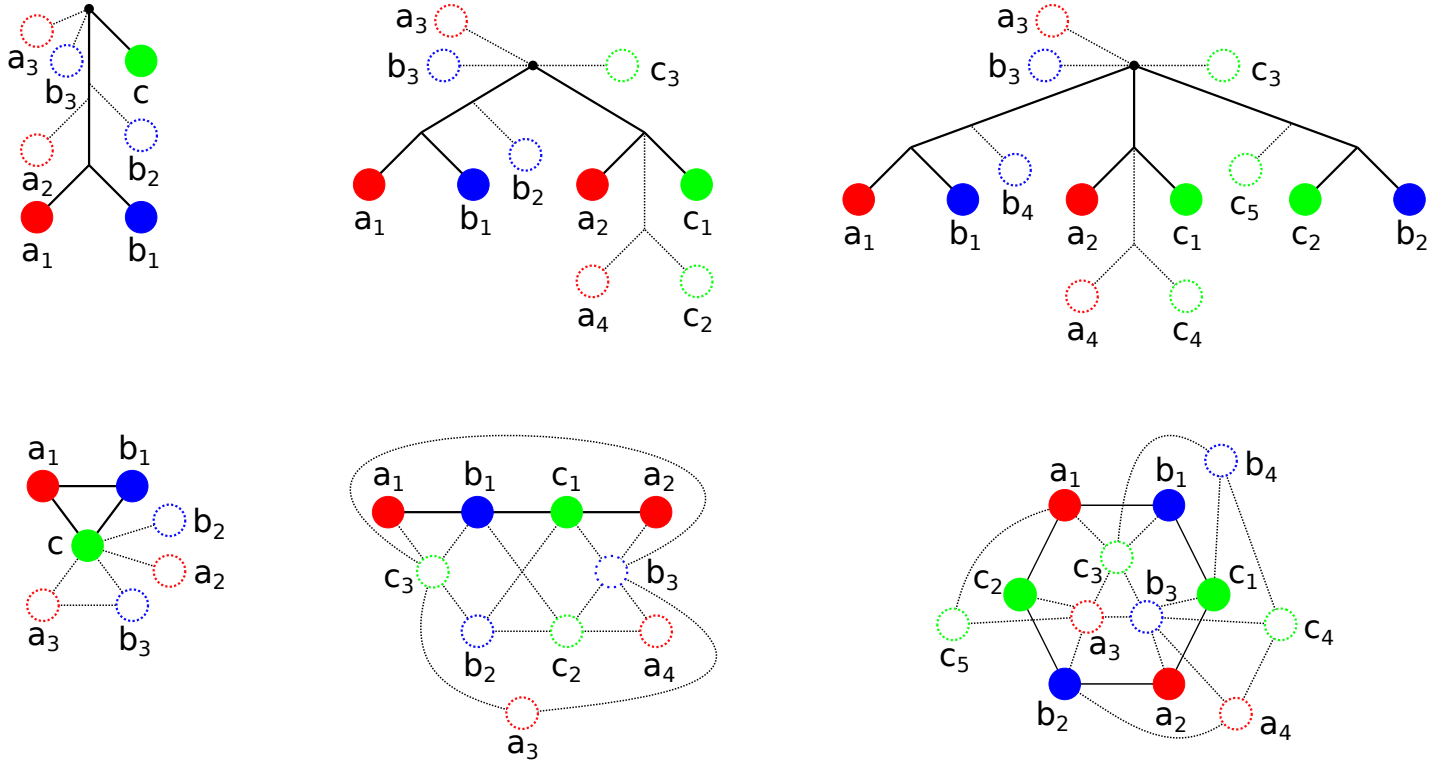
There are exactly three classes of 3-cRBMGs:



(A) (B) (C)

The three classes of 3-cRBMGs

There are exactly three classes of 3-cRBMGs:



(A) (B) (C)

Theorem

A graph (G, σ) is a 3-cRBMG if and only if the construction algorithm returns a tree that explains (G, σ).

What can we say so far about n-cRBMGs?

Idea: Similarly to the case of BMGs, all information needed is contained in the 3-colored induced subgraphs of (G, σ)

Conjecture

An undirected colored graph (G, σ) is an n-cRBMG if and only if for any (G_{rst}, σ_{rst}) there exists a tree (T_{rst}, σ_{rst}) that explains (G_{rst}, σ_{rst}), such that $\mathcal{P} := \bigcup_{r,s,t} T_{rst}$ is compatible.

$(G_{rst}, \sigma_{rst}) :=$ induced subgraph on colors r, s, t of (G, σ)

$\rightarrow$ It looks like there is no polynomial-time construction algorithm for n-cRBMGs