Accurate prediction of orthologs in the presence of divergence after duplication

MANUEL LAFOND, University of Ottawa >> Université de Sherbrooke
MONA M. MIARDAN, University of Ottawa
DAVID SANKOFF, University of Ottawa

In this talk

1. Orthology prediction is **difficult** when ***divergence after duplication*** occurs***.***

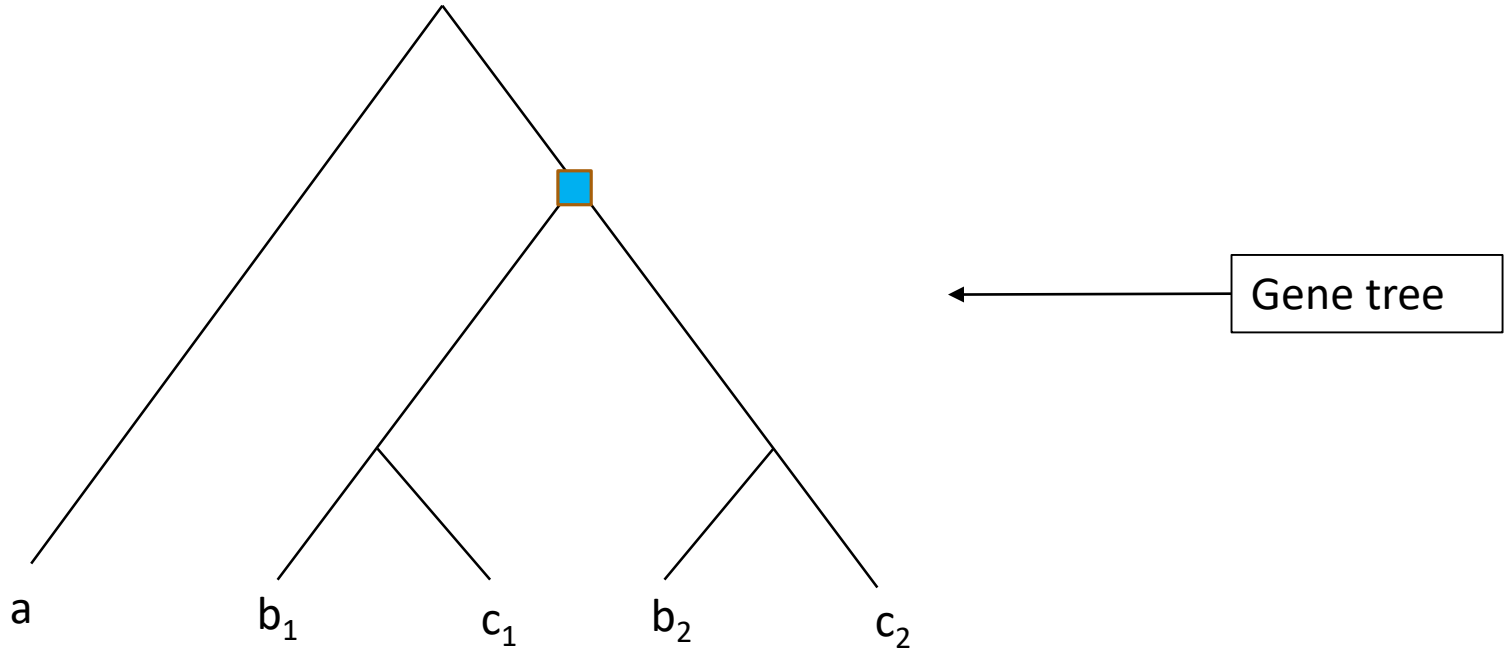
2. An **algorithmic framework** that supports divergence after duplication.

3. **Experiments** on simulated and real datasets.

Let us start with a definition!

Two genes are ***orthologs*** if they descend from an ancestral gene that has undergone **speciation**.

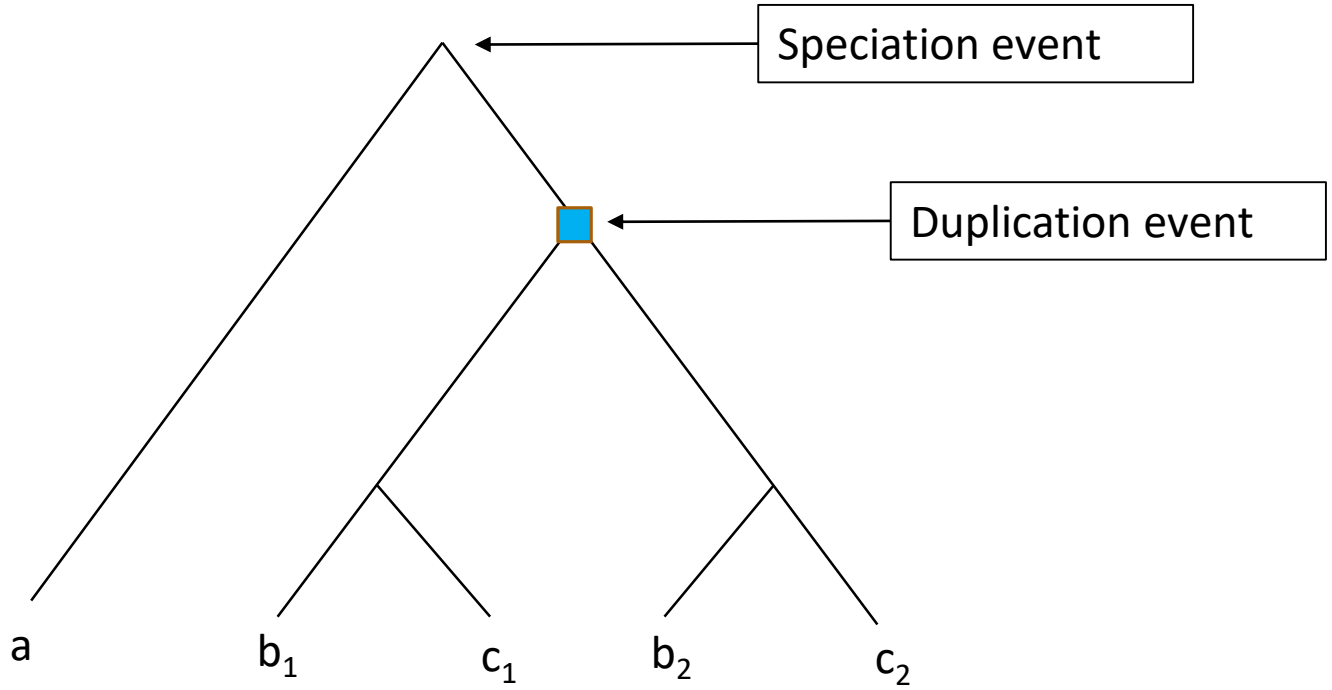
They are ***paralogs*** if they result from **duplication**.



Let us start with a definition!

Two genes are ***orthologs*** if they descend from an ancestral gene that has undergone **speciation**.

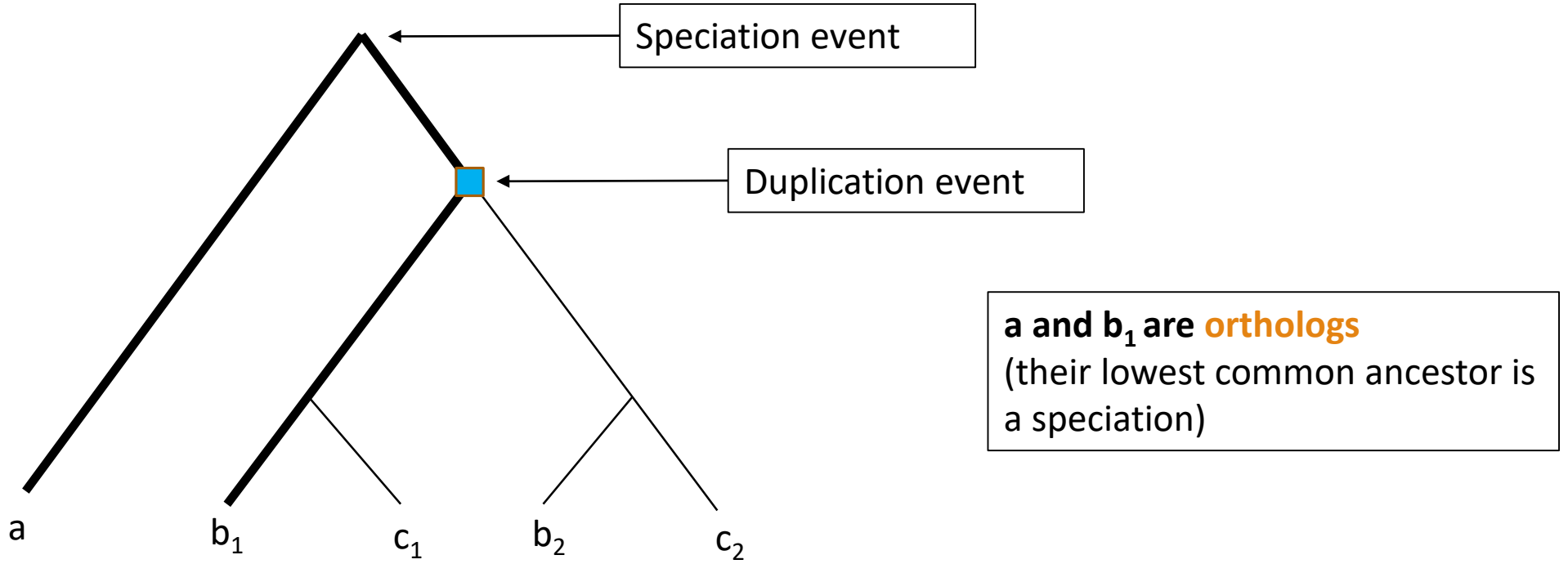
They are ***paralogs*** if they result from **duplication**.



Let us start with a definition!

Two genes are ***orthologs*** if they descend from an ancestral gene that has undergone **speciation**.

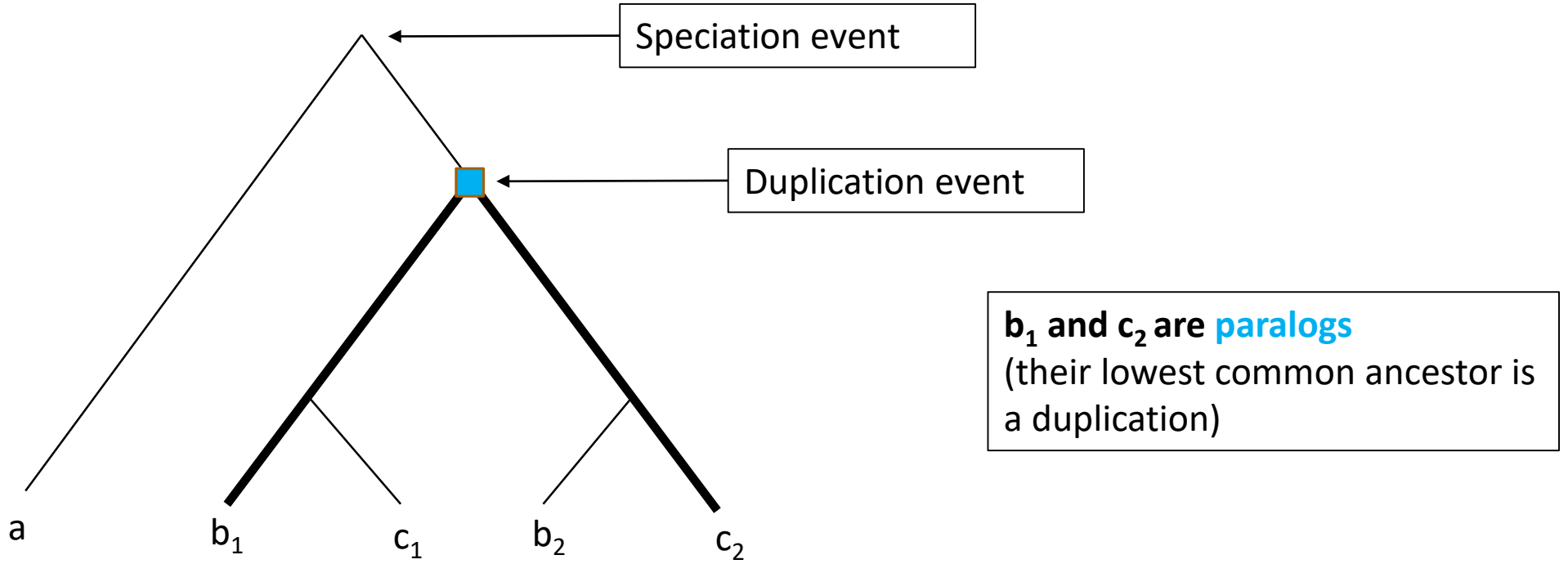
They are ***paralogs*** if they result from **duplication**.



Let us start with a definition!

Two genes are ***orthologs*** if they descend from an ancestral gene that has undergone **speciation**.

They are ***paralogs*** if they result from **duplication**.



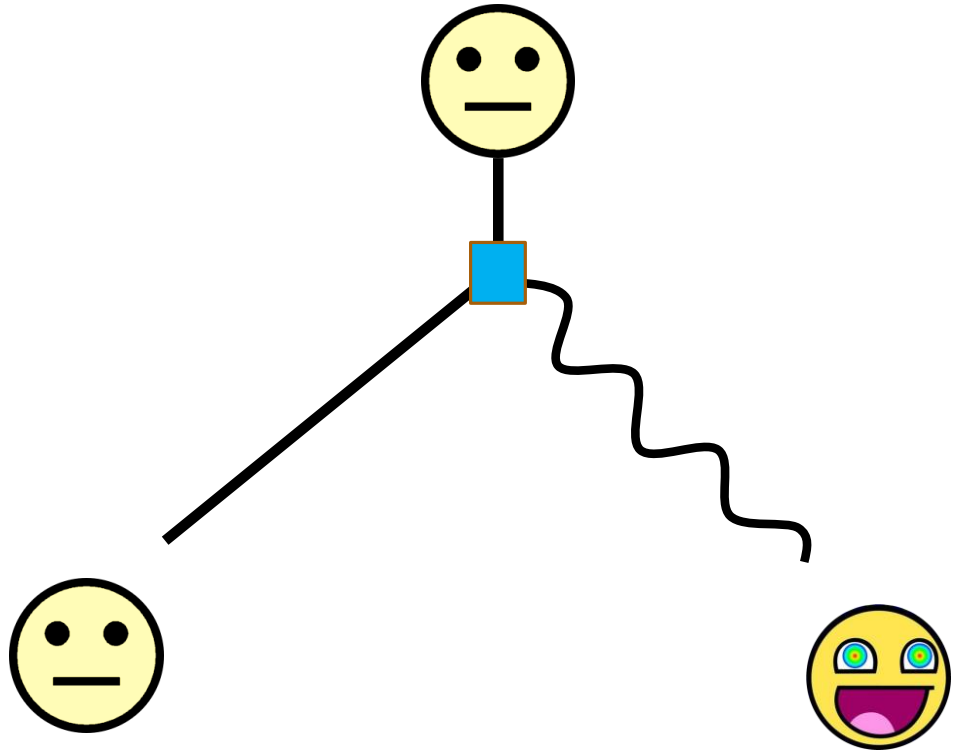
So what???

Orthologs + paralogs are useful to reconstruct **gene trees** + **species trees**.

Orthologous genes can be used to ***predict gene functionality***.

Orthologs conjecture:

- **Orthologs** *tend* to perform ***similar functions*** and share similar DNA sequences.
- **Paralogs** tend to ***diverge*** from the point of view of functions and DNA.



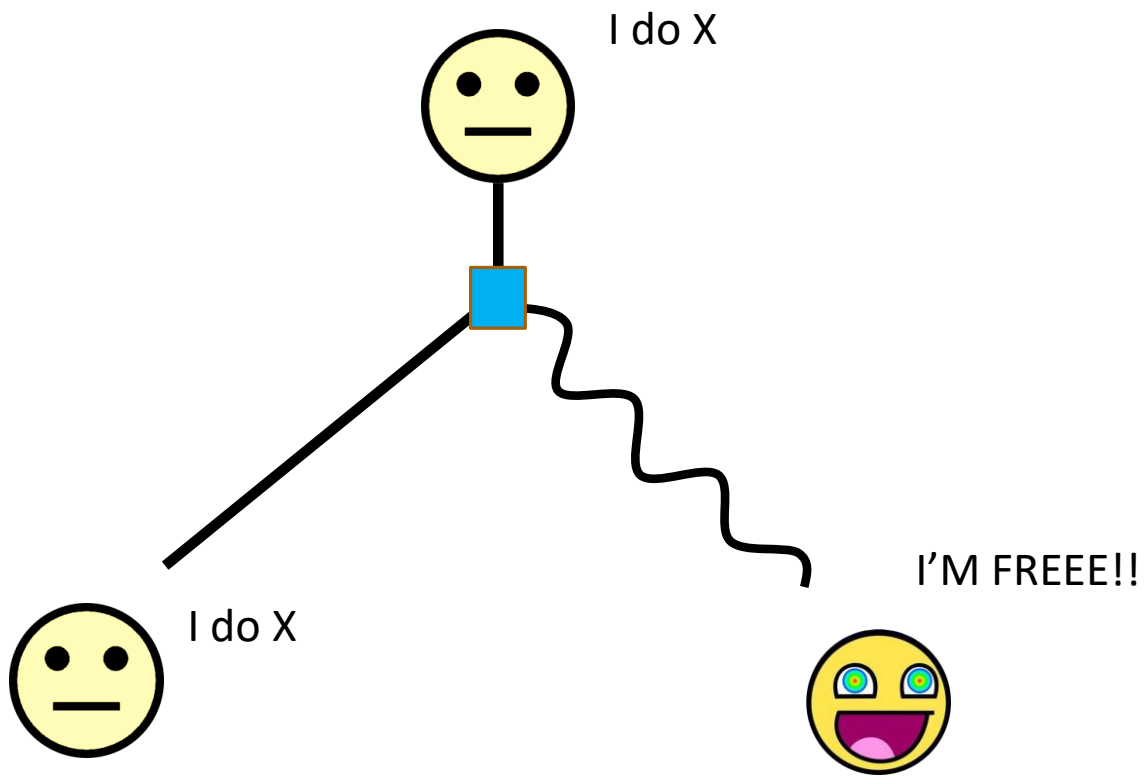
So what???

Orthologs + paralogs are useful to reconstruct **gene trees** + **species trees**.

Orthologous genes can be used to ***predict gene functionality***.

Orthologs conjecture:

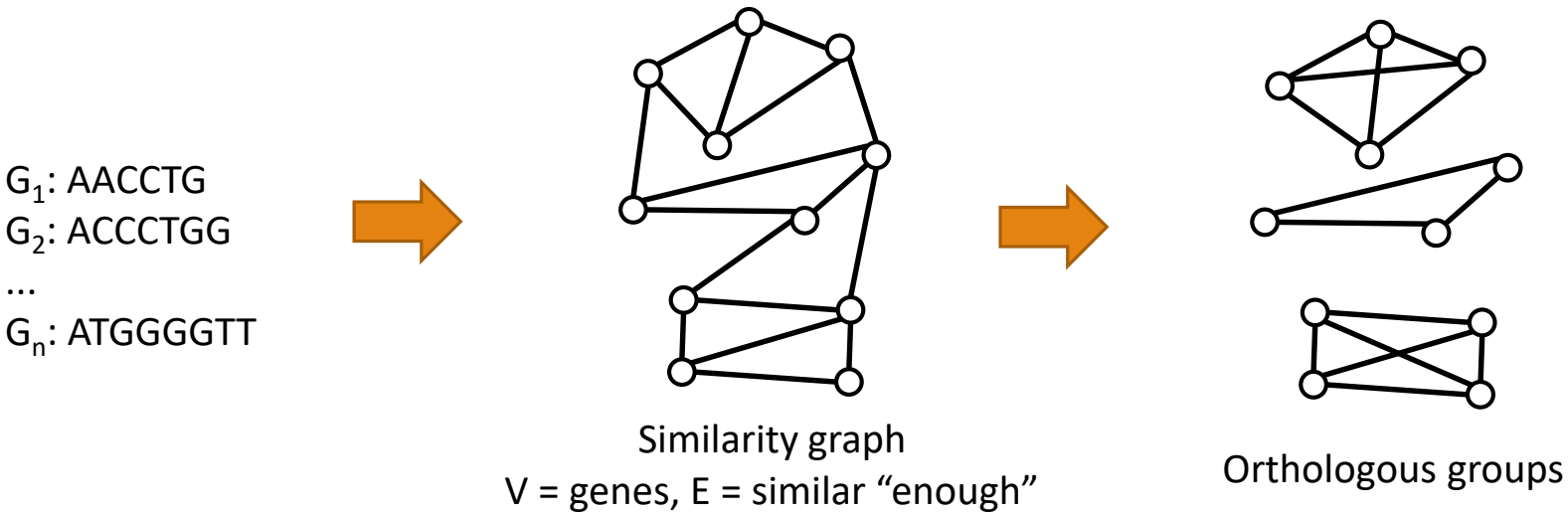
- **Orthologs** *tend* to perform ***similar functions*** and share similar DNA sequences.
- **Paralogs** tend to ***diverge*** from the point of view of functions and DNA.



And how do we find orthologs?

Similarity-based methods:

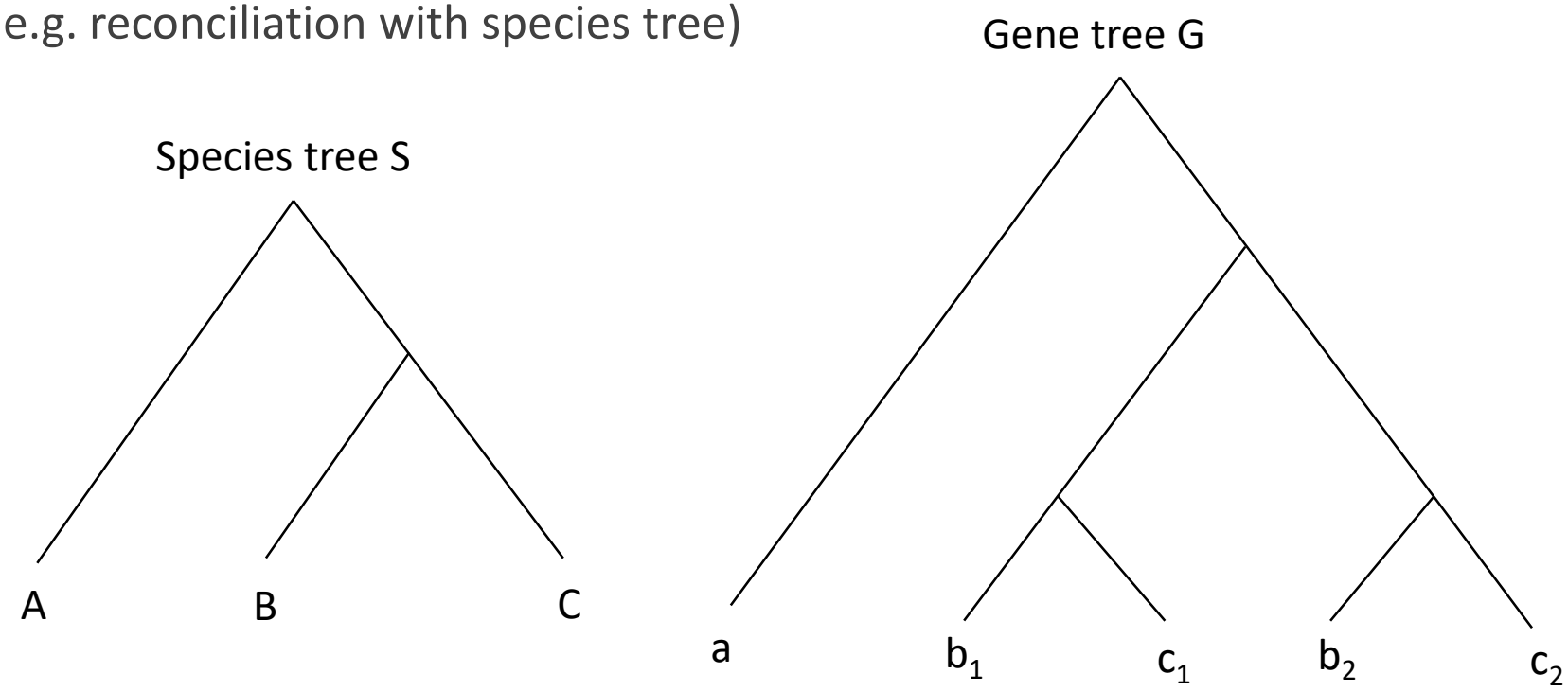
- Assume DNA similarity => Orthology.
- Build a *similarity graph* (edge weights = similarity).
- Cluster the genes into orthologous groups.



And how do we find orthologs?

Phylogeny-based methods:

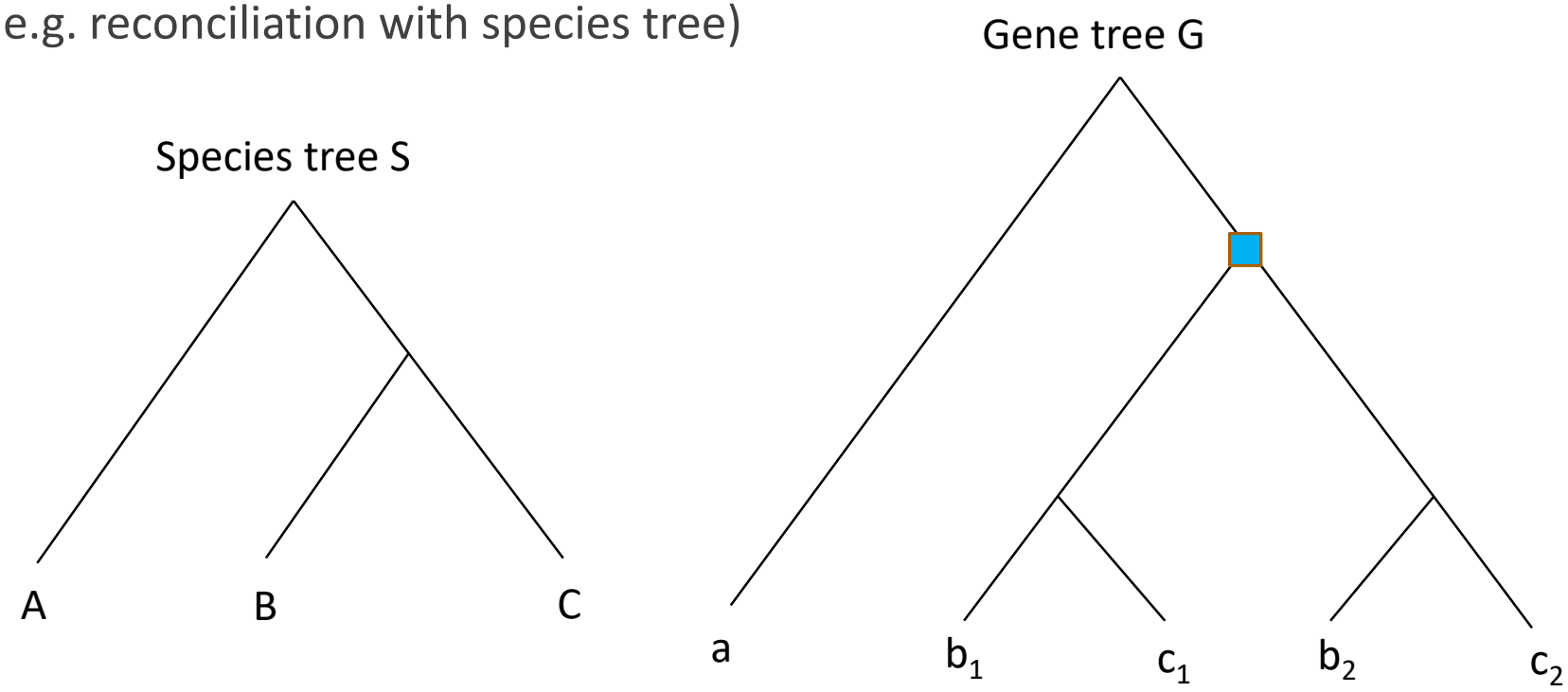
- Build a gene tree.
- Identify speciation and duplication events.
  - (using e.g. reconciliation with species tree)



And how do we find orthologs?

Phylogeny-based methods:

- Build a gene tree.
- Identify speciation and duplication events.
  - (using e.g. reconciliation with species tree)



Some related work (non-exhaustive)

Similarity-based methods

- OrthoMCL
- ProteinOrtho
- OMA / OMA-GETHOGS
- OrthoFinder
- COG, eggNOG, InParanoid, and more ...

Phylogeny-based methods

- LOFT
- COCO-CL
- Reconciliation tools (e.g. NOTUNG)
- Cograph editing tools (Hellmuth group, L, El-Mabrouk, Dondi, ...)

*sorry if you think I should've mentioned your work here – if so, let me know!

Divergence after duplication

After duplication, asymmetric rates of evolution may occur.

- *Standard model*: one copy keeps ***same rate***, the other acquires ***accelerated rate***.

Category I								
Neofunc-tionalization	Kept	Novel	Gain-of-function mutations	Neutral	Purifying selection	Neutral	α	β
DDC	Subfunc-tionalized	Subfunc-tionalized	Loss-of-function mutations	Neutral	Relaxed purifying selection	Relaxed purifying selection	β	β
Specialization or EAC	Subfunc-tionalized	Subfunc-tionalized	Gain-of-function mutations	Neutral	Relaxed purifying selection	Relaxed purifying selection	β	β
Category II								
Positive dosage	Kept	Same as original	NA	Positive selection on duplication	NA	NA	α′	α′
Shielding against deleterious mutations	Kept	Same as original	NA	Positive selection on duplication	Relaxed purifying selection	Relaxed purifying selection	NA	NA
Modified duplication	Kept	Novel	Gain-of-function mutations	Positive selection on duplication	NA	NA	α	β
Category III								
Permanent heterozygote	Subfunc-tionalized	Subfunc-tionalized	Gain-of-function mutations	Positive selection on pre-duplicational variation	NA	NA	β	β
Adaptive radiation model	Kept	Novel	Gain-of-function mutations	Positive selection on pre-duplicational variation	NA	NA	α	β
Diversifying selection	Multiple functions	Multiple functions	Gain-of-function mutations	Positive selection on pre-duplicational variation	NA	NA	o	o
Category IV								
Dosage balance	Kept	Original	NA	NA	NA	NA	α′	α′

α = same rate
β = faster rate

Innan, Hideki, and Fyodor Kondrashov. "The evolution of gene duplications: classifying and distinguishing between models." Nature Reviews Genetics (2010).

Divergence after duplication

After duplication, asymmetric rates of evolution may occur.

- *Standard model*: one copy keeps ***same rate***, the other acquires ***accelerated rate***.

Divergence after duplication

After duplication, asymmetric rates of evolution may occur.

- *Standard model*: one copy keeps ***same rate***, the other acquires ***accelerated rate***.
- Not always the case, but *it does happen*.

Divergence after duplication

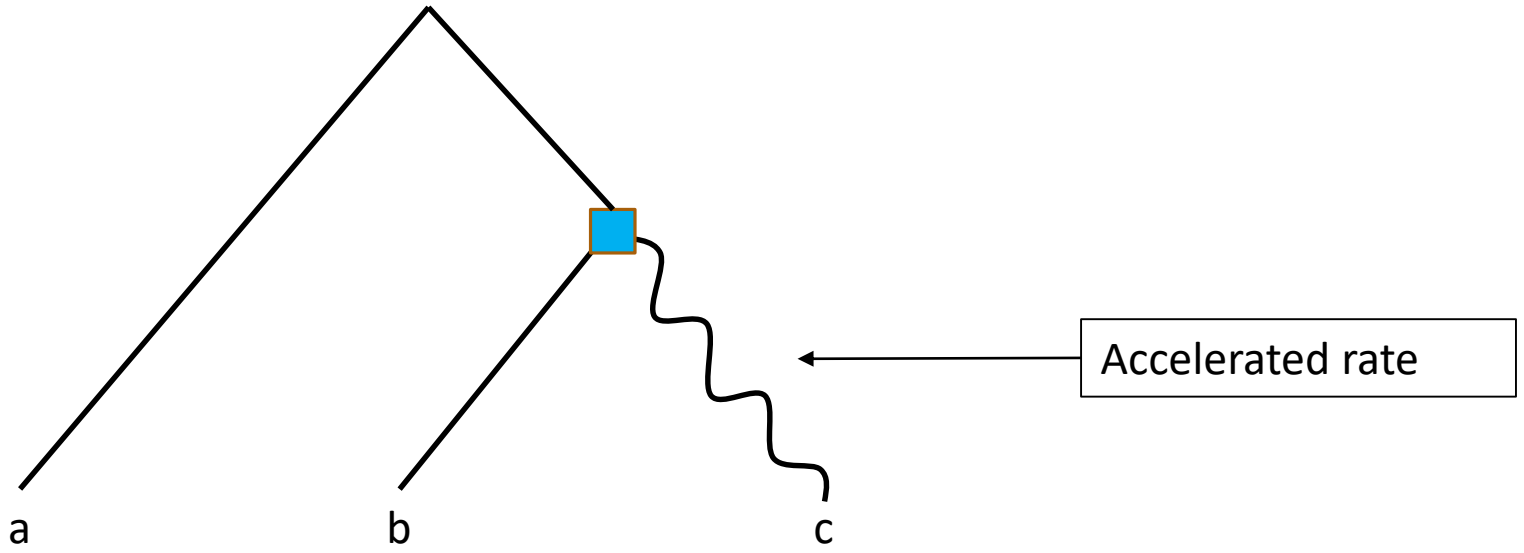
After duplication, asymmetric rates of evolution may occur.

- *Standard model*: one copy keeps ***same rate***, the other acquires ***accelerated rate***.
- Not always the case, but *it does happen*.

- When it does, how well do orthology prediction methods fare?

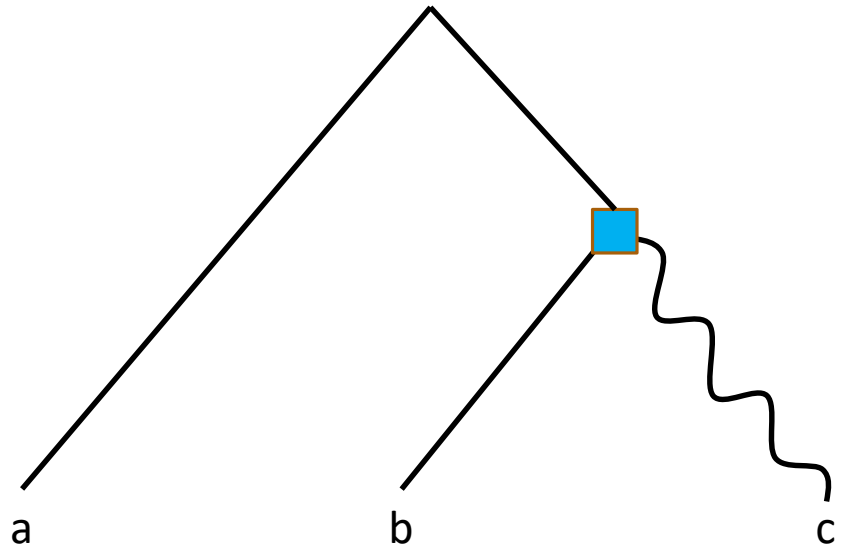
Divergence after duplication

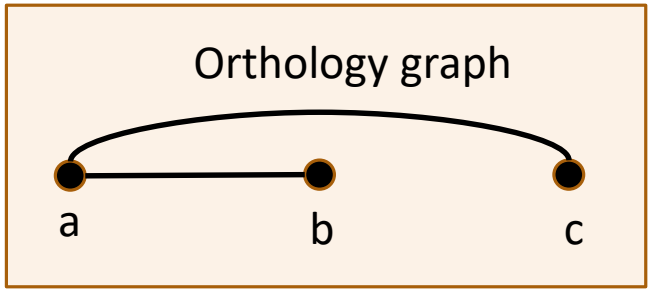
Suppose this is the true history.



Divergence after duplication

Suppose this is the true history.

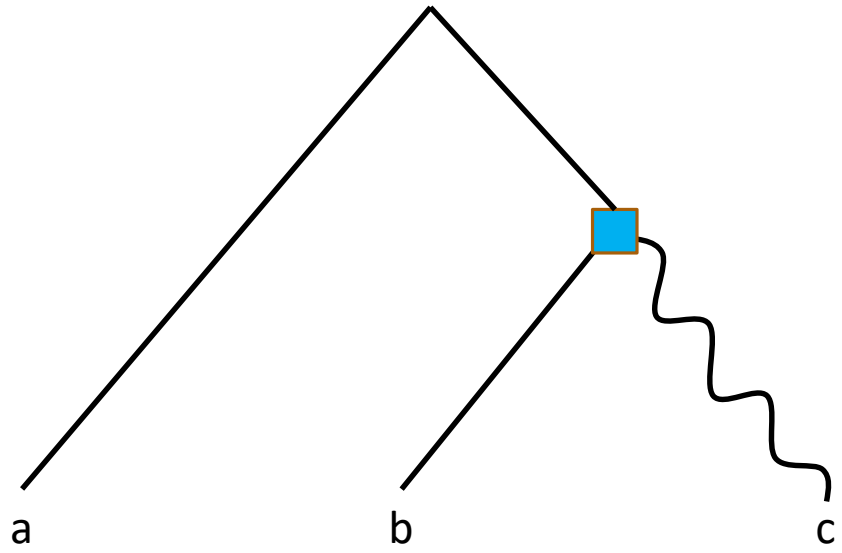


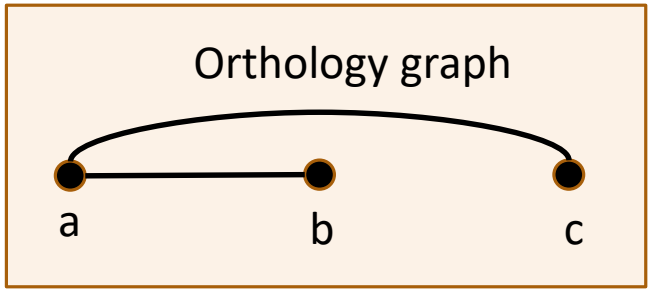


Divergence after duplication

Suppose this is the true history.

- (a,b) are “similar”, but (a,c) and (b, c) are “not similar”.

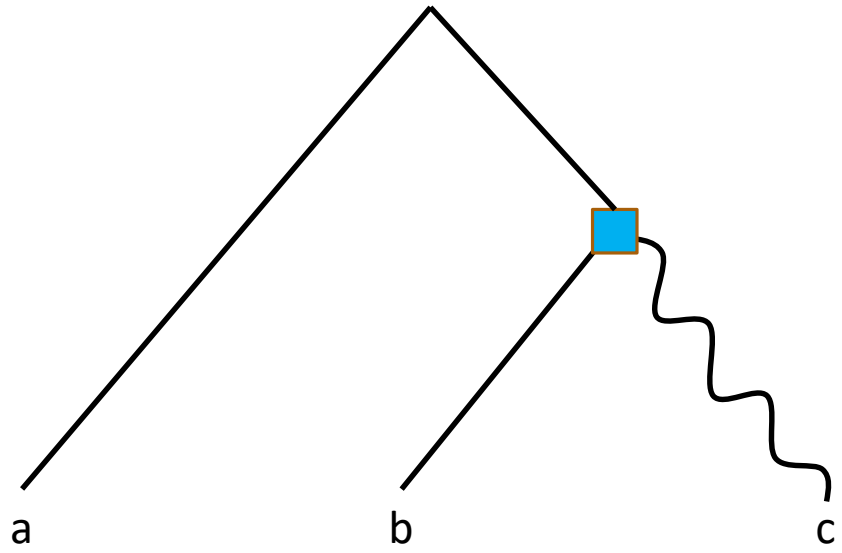


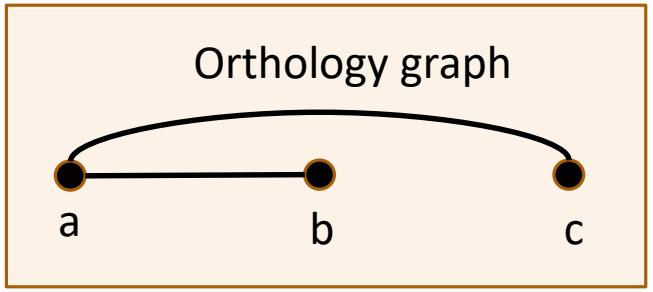


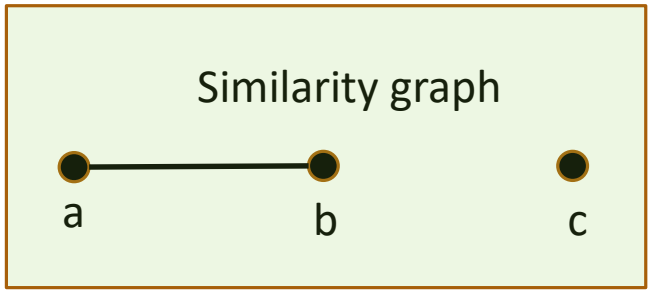
Divergence after duplication

Suppose this is the true history.

- (a,b) are “similar”, but (a,c) and (b, c) are “not similar”.

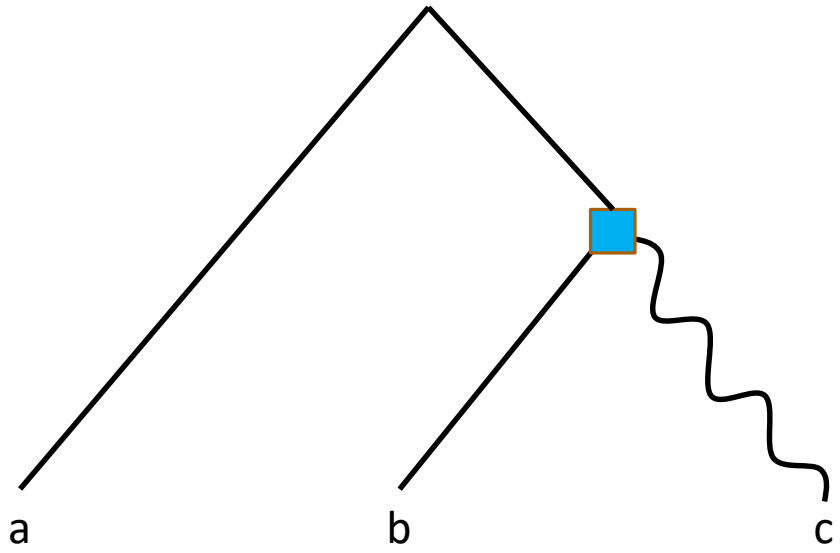


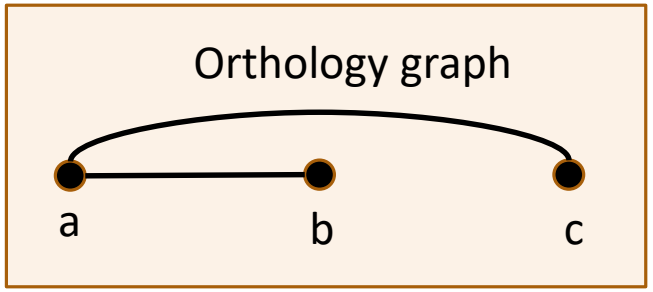


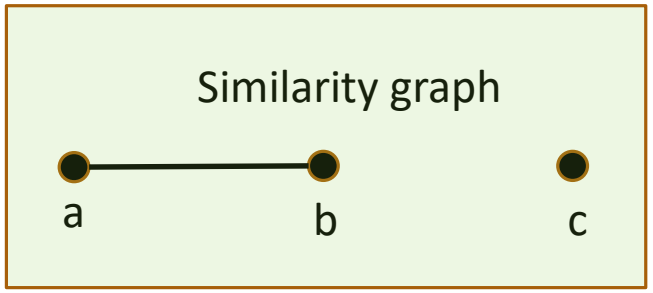


Divergence after duplication

Similarity-based methods might miss “non-similar” orthologs.





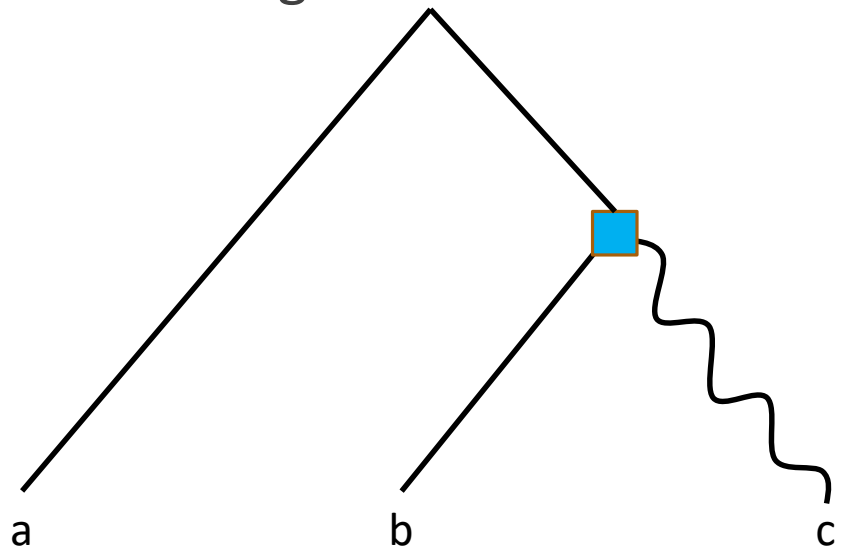


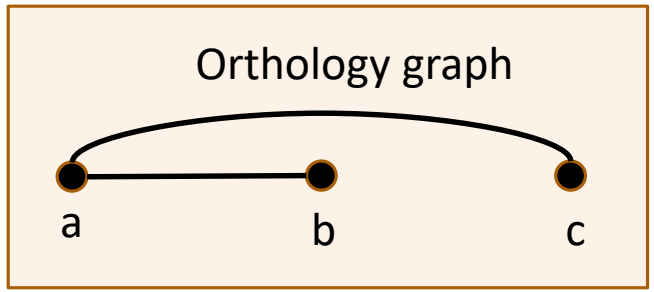
= orthologs predicted by similarity-based methods

Divergence after duplication

Phylogeny-based methods:

- will find all orthologs (if we are lucky)...
- but **do not distinguish** “similar” orthologs and “non-similar” orthologs.
- no good for functional analysis studies.

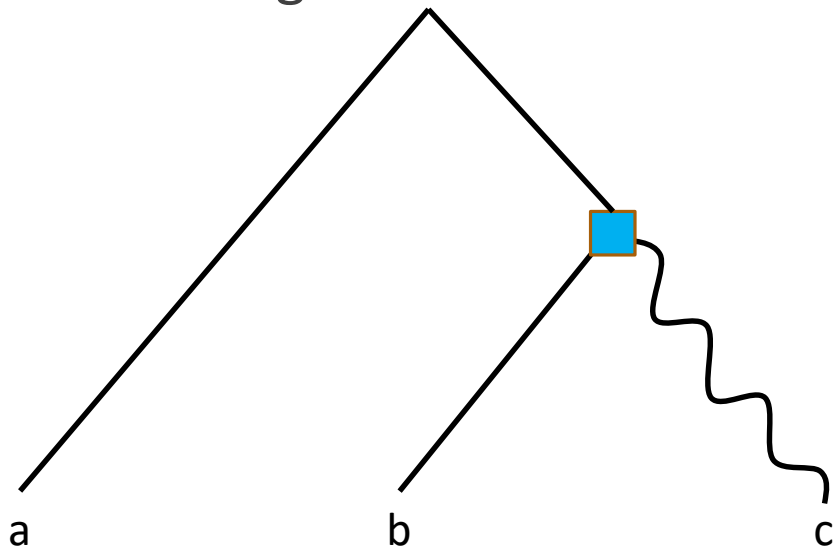


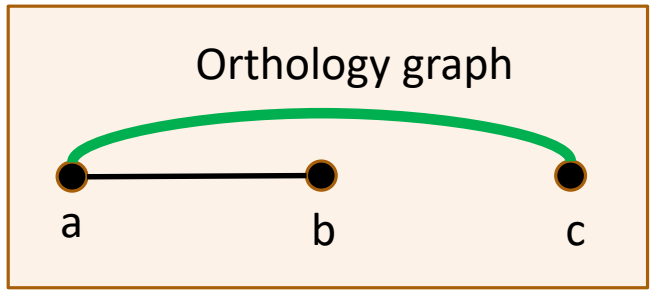


Divergence after duplication

Phylogeny-based methods:

- will find all orthologs (if we are lucky)...
- but **do not distinguish** “similar” orthologs and “non-similar” orthologs.
- no good for functional analysis studies.





Objectives of this work

1. Find **all orthologs** even in the presence of **divergence after duplication**

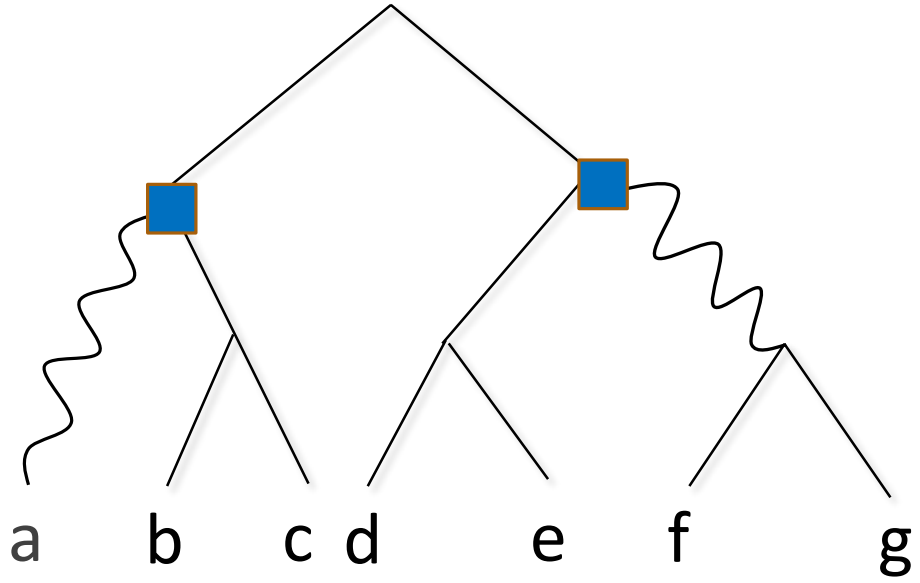
2. Distinguish between **“similar”** orthologs, and **“non-similar”** orthologs.
   - Hereafter called ***primary*** and ***secondary*** orthologs, respectively.
   - Primary orthologs are potential ***isoorthologs*** and potential ***equivalogs***.
     - ***Isoorthologs :*** *orthologs that have retained their function up to their lowest common ancestor*
     - ***Equivalogs :*** *same definition, but not limited to orthologs*

Algorithmic framework

DAD model (DAD = Divergence After Dup)

Assume that **every** duplication introduces exactly one **divergent edge**.

- (yes, a bit extreme I know)



Call two genes **primary orthologs** if

- they are orthologs;
- there is no divergent edge on their unique path in the gene tree.

Call two genes **secondary orthologs** if they are orthologs but not primary.

DAD model (DAD = Divergence After Dup)

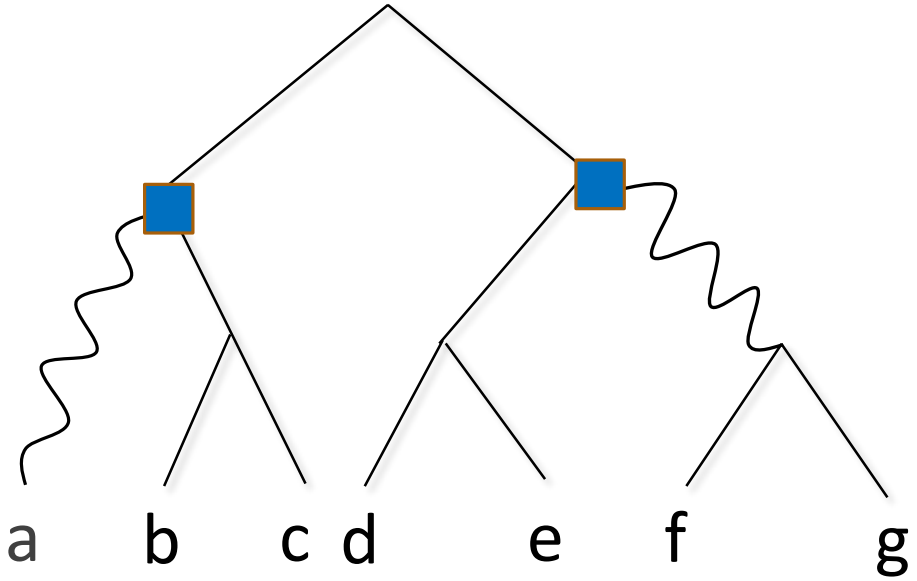
Assume that **every** duplication introduces exactly one **divergent edge**.

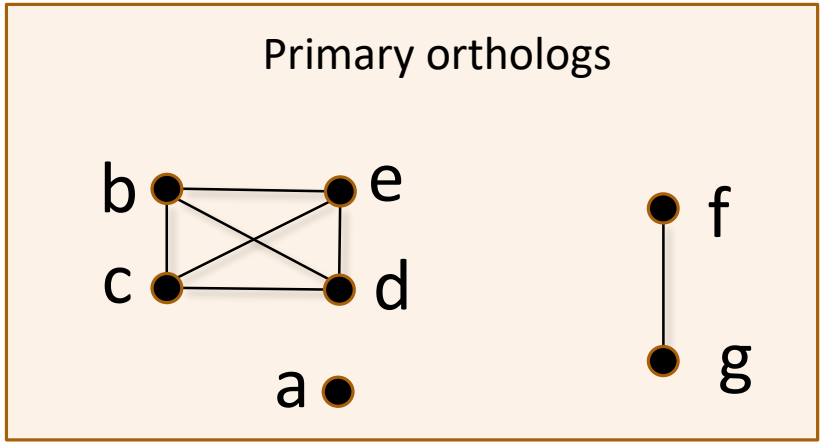
- (yes, a bit extreme I know)

Call two genes **primary orthologs** if

- they are orthologs;
- there is no divergent edge on their unique path in the gene tree.

Call two genes **secondary orthologs** if they are orthologs but not primary.

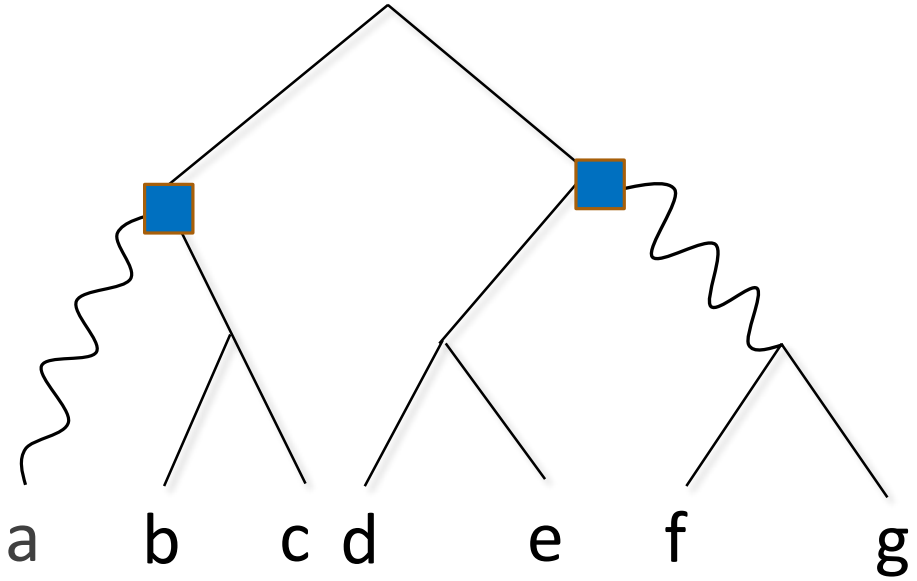




DAD model (DAD = Divergence After Dup)

Assume that **every** duplication introduces exactly one **divergent edge**.

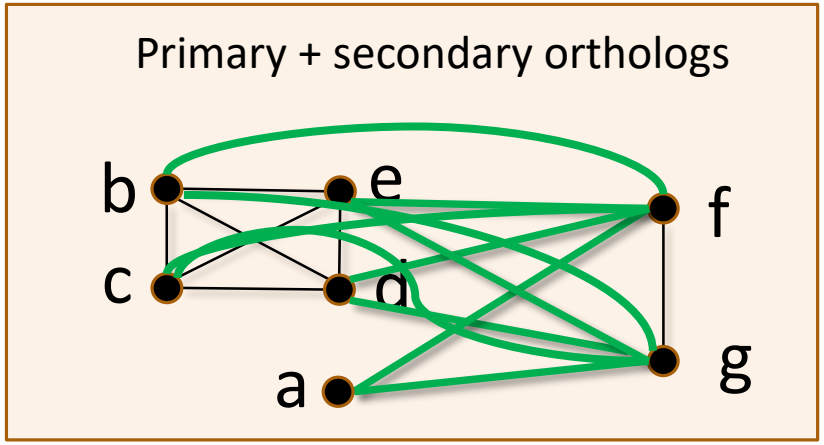
- (yes, a bit extreme I know)



Call two genes **primary orthologs** if

- they are orthologs;
- there is no divergent edge on their unique path in the gene tree.

Call two genes **secondary orthologs** if they are orthologs but not primary.



Characterizing primary orthologs

Proposition

Under the DAD model, **primary orthology** is an **equivalence relation**

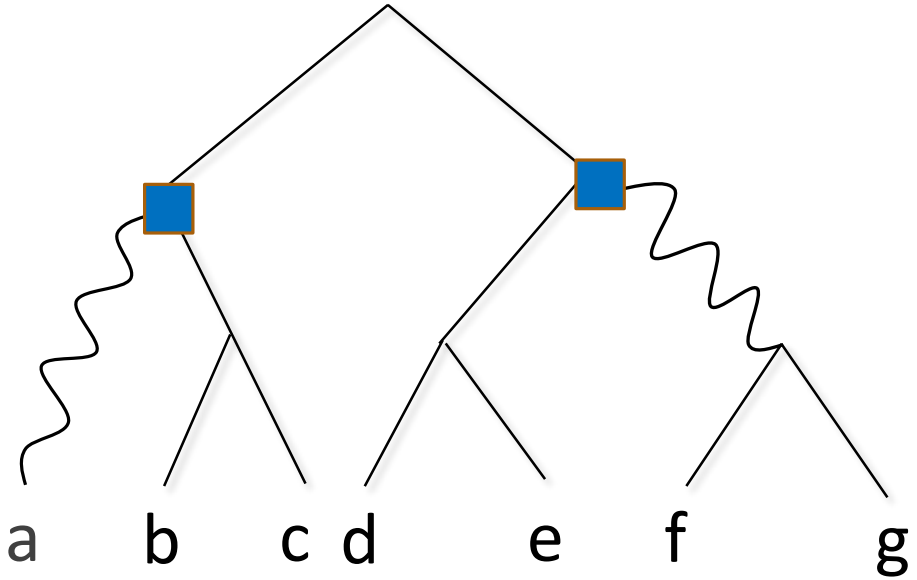
(i.e. primary orthologs form *a collection of disjoint cliques*).

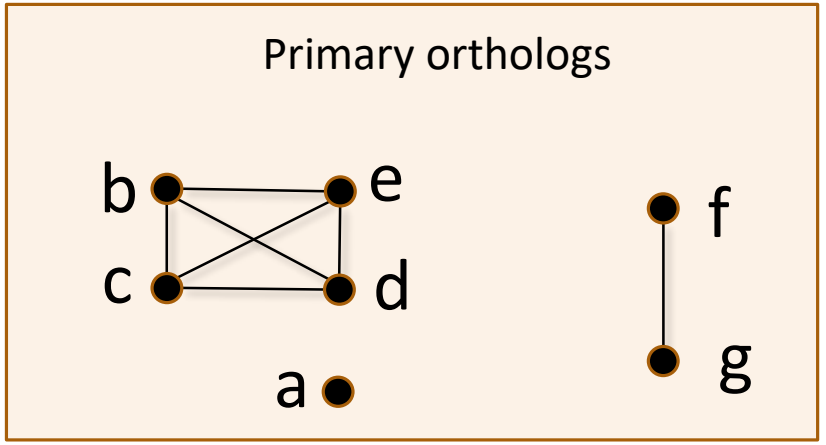
Characterizing primary orthologs

Proposition

Under the DAD model, **primary orthology** is an **equivalence relation**

(i.e. primary orthologs form *a collection of disjoint cliques*).



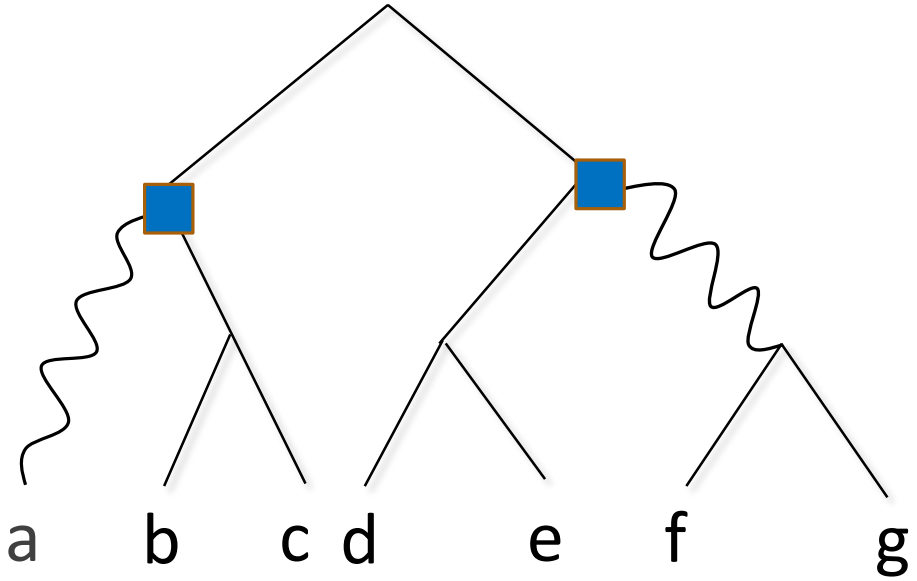


Characterizing primary orthologs

Proposition

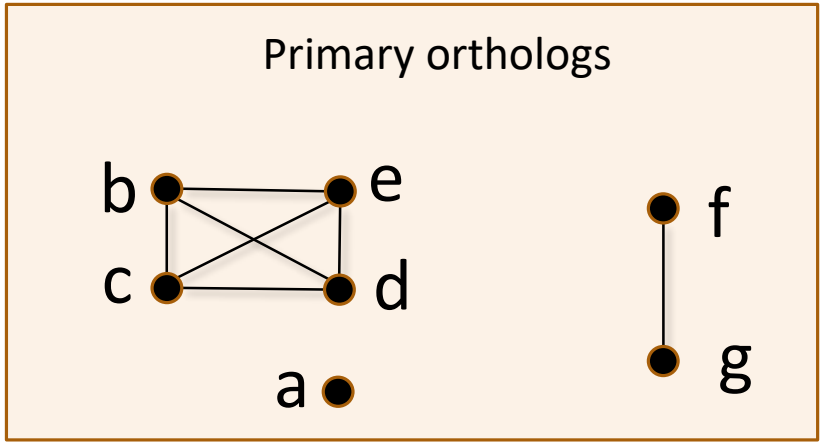
Under the DAD model, **primary orthology** is an **equivalence relation**

(i.e. primary orthologs form *a collection of disjoint cliques*).



"Corollary"

Similarity-based methods that find 1-to-1 orthologs find the primary orthologs.



HyPPO framework

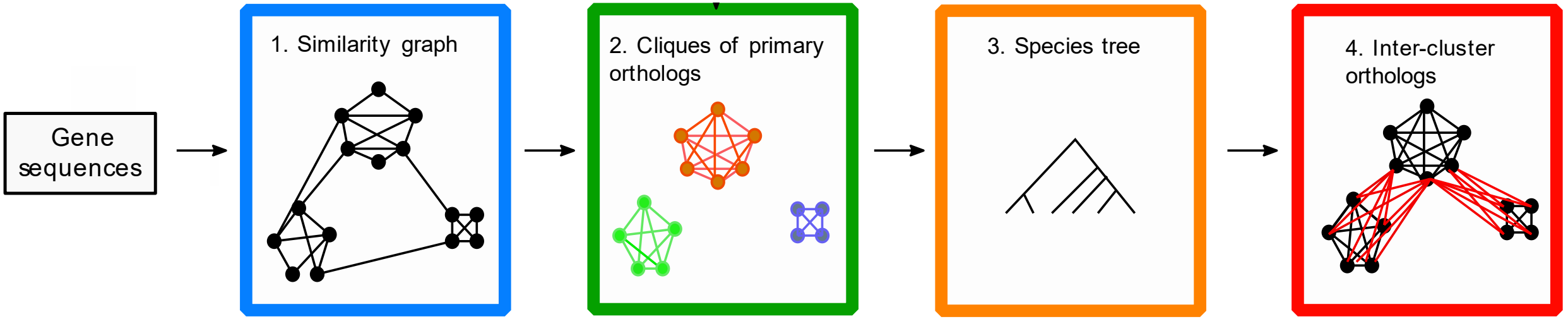
HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs

Given a gene family:

1. Calculate **scores** between each gene pair based on DNA.
2. Compute the **cliques of primary orthologs**.
3. Infer the **inter-cluster orthologs** (the secondary orthologs).
   - If we are not careful, may result in **non-sensical** orthologs.
   - A **species tree** is needed for this step (why? see paper).

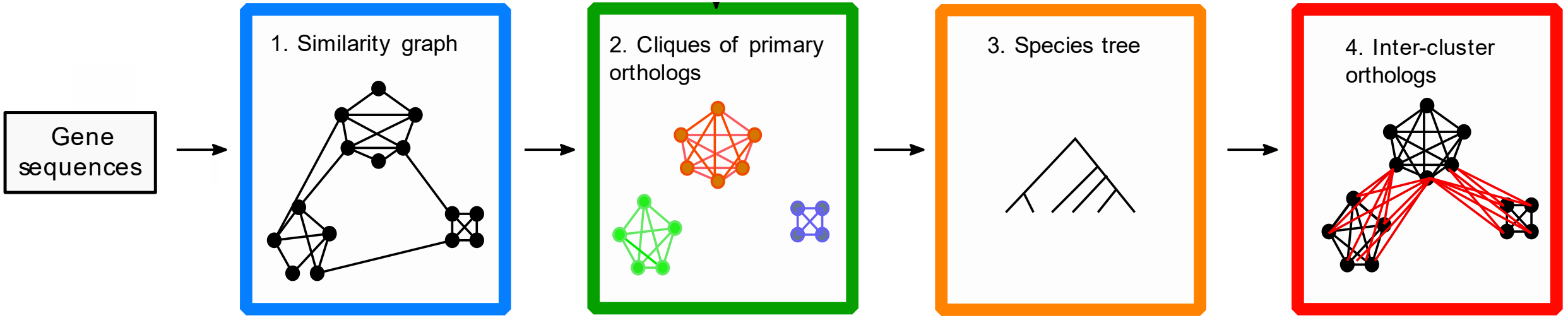
The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs



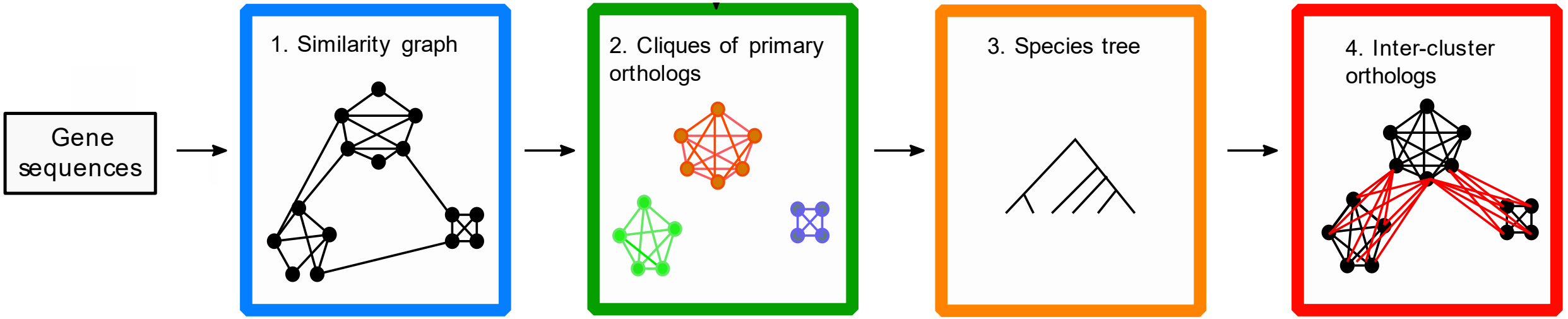
The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs
Steps 2-3-4 define three algorithmic problems.



The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs
Steps 2-3-4 define three algorithmic problems.



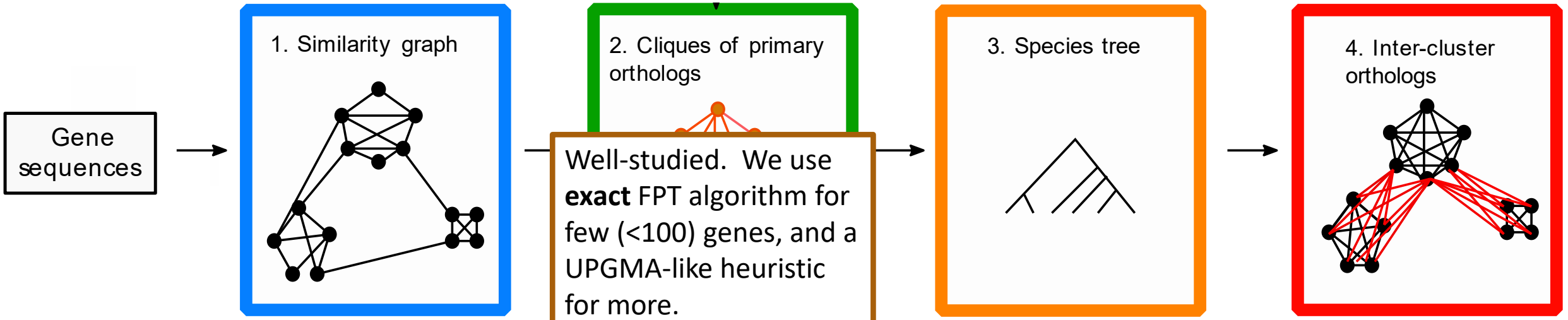
CLUSTER EDITING PROBLEM
Edit min # of edges to get collection of cliques.

SPECIES TREE FROM CLUSTERS PROBLEM
Find species tree S that "agrees" with most clusters.

MAX CONSISTENT ORTHOLOGY
Find max weight orthologs that are consistent with S.

The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs
Steps 2-3-4 define three algorithmic problems.



Well-studied. We use **exact** FPT algorithm for few (<100) genes, and a UPGMA-like heuristic for more.

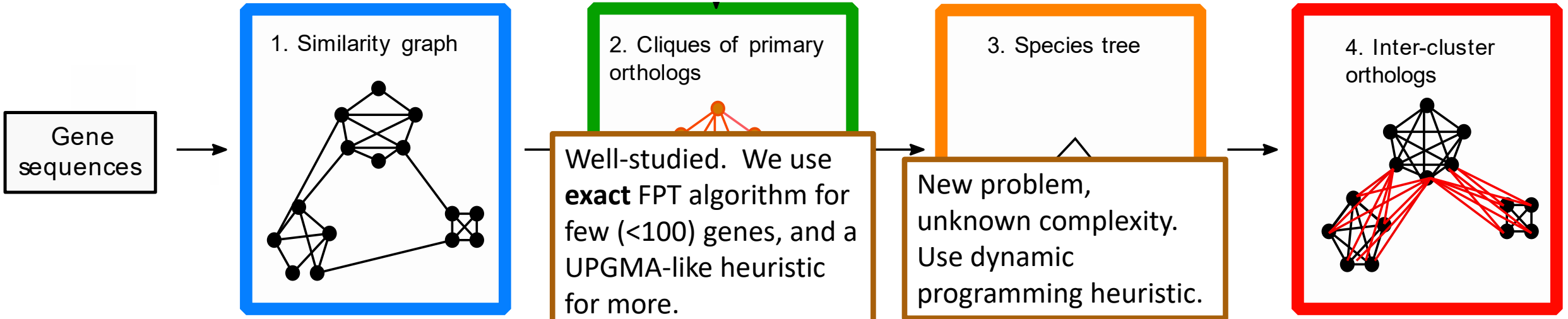
CLUSTER EDITING PROBLEM
Edit min # of edges to get collection of cliques.

SPECIES TREE FROM CLUSTERS PROBLEM
Find species tree S that "agrees" with most clusters.

MAX CONSISTENT ORTHOLOGY
Find max weight orthologs that are consistent with S.

The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs
Steps 2-3-4 define three algorithmic problems.



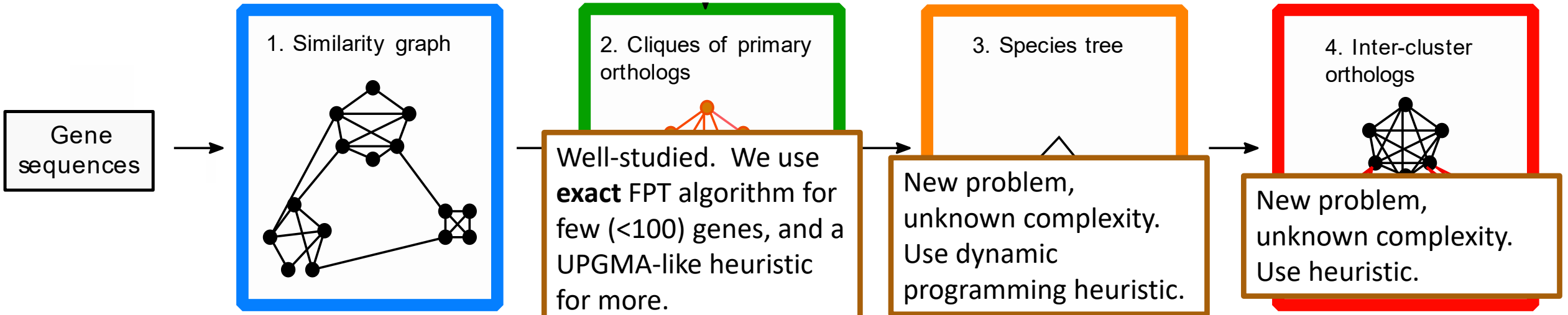
CLUSTER EDITING PROBLEM
Edit min # of edges to get collection of cliques.

SPECIES TREE FROM CLUSTERS PROBLEM
Find species tree S that "agrees" with most clusters.

MAX CONSISTENT ORTHOLOGY
Find max weight orthologs that are consistent with S.

The HyPPO framework

HyPPO = **Hy**brid **P**rediction of **P**aralogs and **O**rthologs
Steps 2-3-4 define three algorithmic problems.



CLUSTER EDITING PROBLEM
Edit min # of edges to get collection of cliques.

SPECIES TREE FROM CLUSTERS PROBLEM
Find species tree S that "agrees" with most clusters.

MAX CONSISTENT ORTHOLOGY
Find max weight orthologs that are consistent with S.

Experimental results

Simulated datasets

Using ***SimPhy + INDELible***:

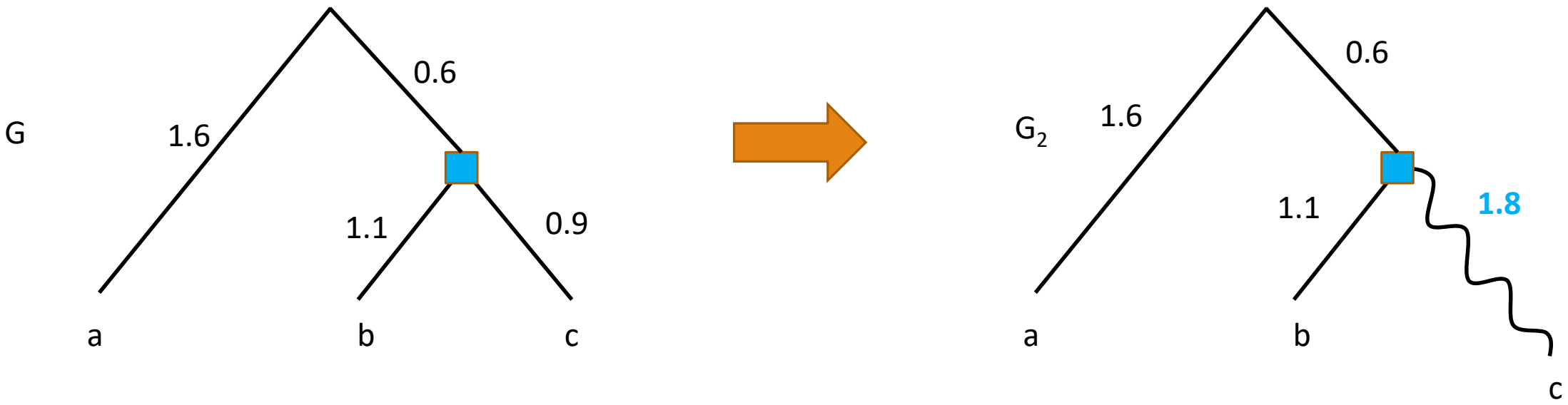
- Generated 40 species trees on [30-50] leaves.
- For each species tree, generated 10 gene families (gene trees).
  - Simulates speciation, duplication and losses (with random nucleotide evolutionary model).
  - Parameters evaluated: **dup/loss rates**, **substitution rate**, **dup-divergence rate**.

Dataset	Dup rate	Loss rate	Substitution rate	Nb trees
Standard	5e-7	5e-7	5e-6	10 species trees, 100 gene trees
Eventful	**1e-7**	**1e-7**	5e-6	10 species trees, 100 gene trees
Fast	5e-7	5e-7	**5e-5**	10 species trees, 100 gene trees
Slow	5e-7	5e-7	**5e-7**	10 species trees, 100 gene trees

Simulating divergence after duplication

For each gene tree G, generate a gene tree G_2 by:

- choosing one divergent edge ***e*** per duplication
- multiplying the length of ***e*** by 2

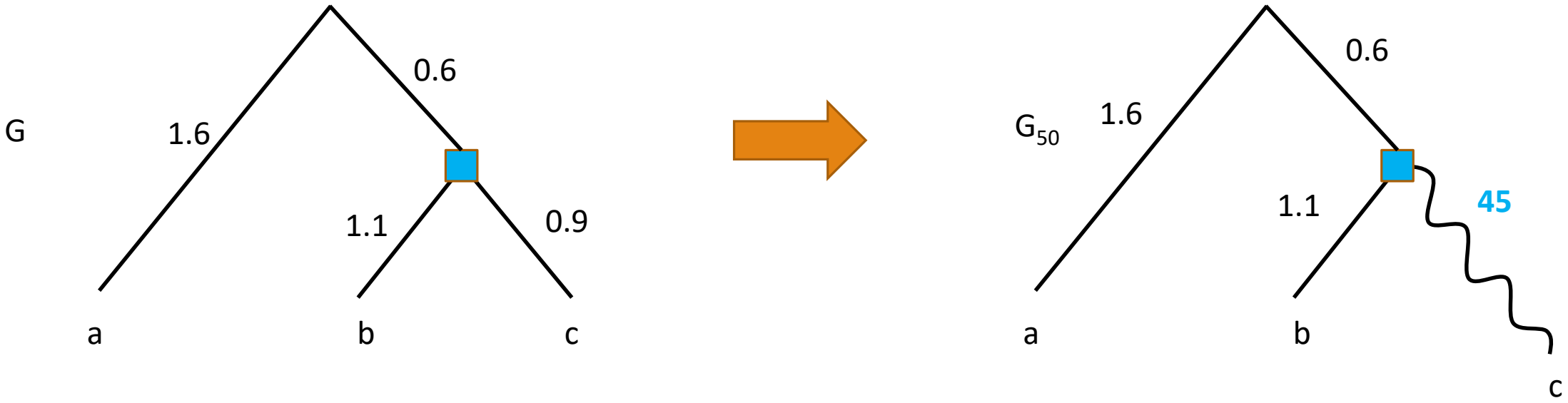


Simulating divergence after duplication

For each gene tree G, generate a gene tree G_2 by:

- choosing one divergent edge ***e*** per duplication
- multiplying the length of ***e*** by 2

Did the same for G_8 and G_{50}



Dataset	Dup rate	Loss rate	Substitution rate	Nb trees
Standard	5e-7	5e-7	5e-6	10 species trees, 100 gene trees x1 100 gene trees x2 100 gene trees x8 100 gene trees x50
Eventful	**1e-7**	**1e-7**	5e-6	10 species trees, 100 gene trees x1 100 gene trees x2 100 gene trees x8 100 gene trees x50
Fast	5e-7	5e-7	**5e-5**	10 species trees, 100 gene trees x1 100 gene trees x2 100 gene trees x8 100 gene trees x50
Slow	5e-7	5e-7	**5e-7**	10 species trees, 100 gene trees x1 100 gene trees x2 100 gene trees x8 100 gene trees x50

Methods considered

1. **HyPPO**: our whole pipeline

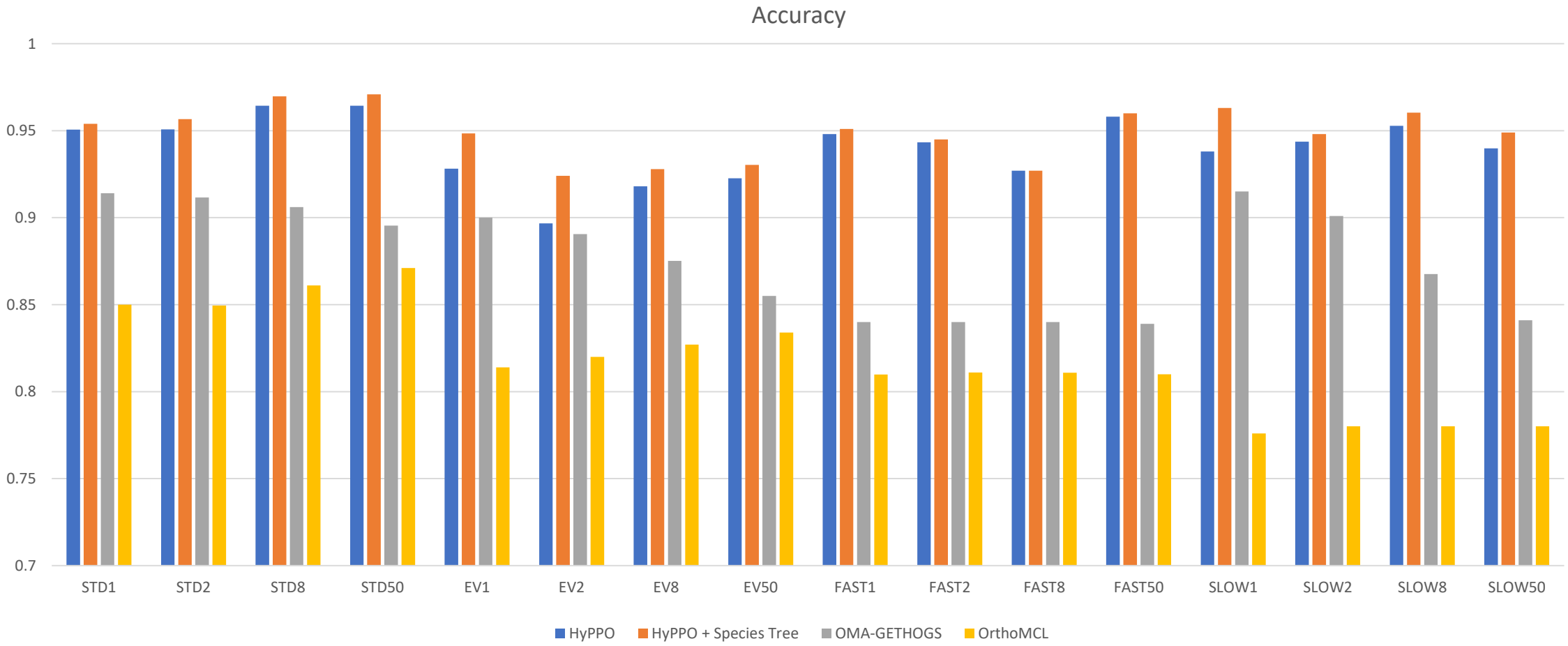
2. **HyPPO + species tree**: same, but true species tree known

3. **OMA-GETHOGS**: can predict primary orthologs.

4. **OrthoMCL**

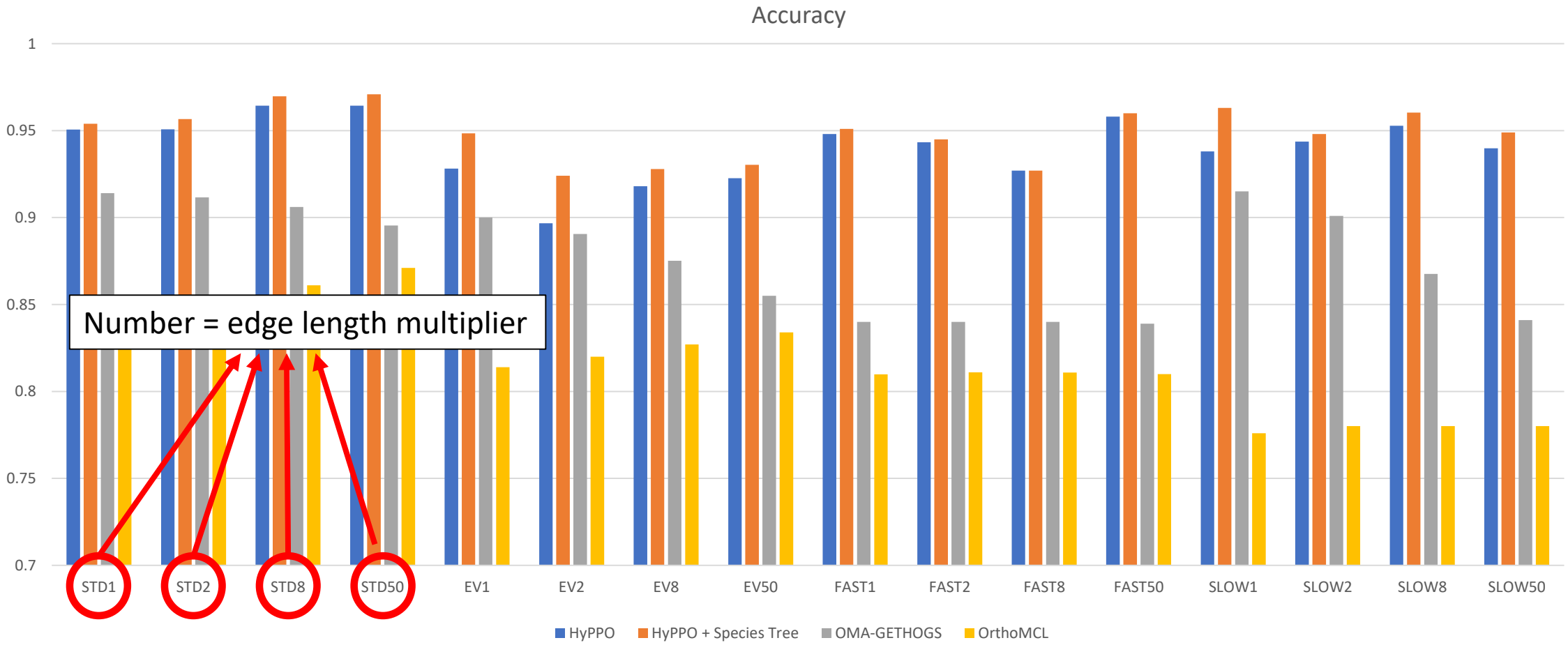
Accuracy on simulated trees

of correct relations / # of gene pairs



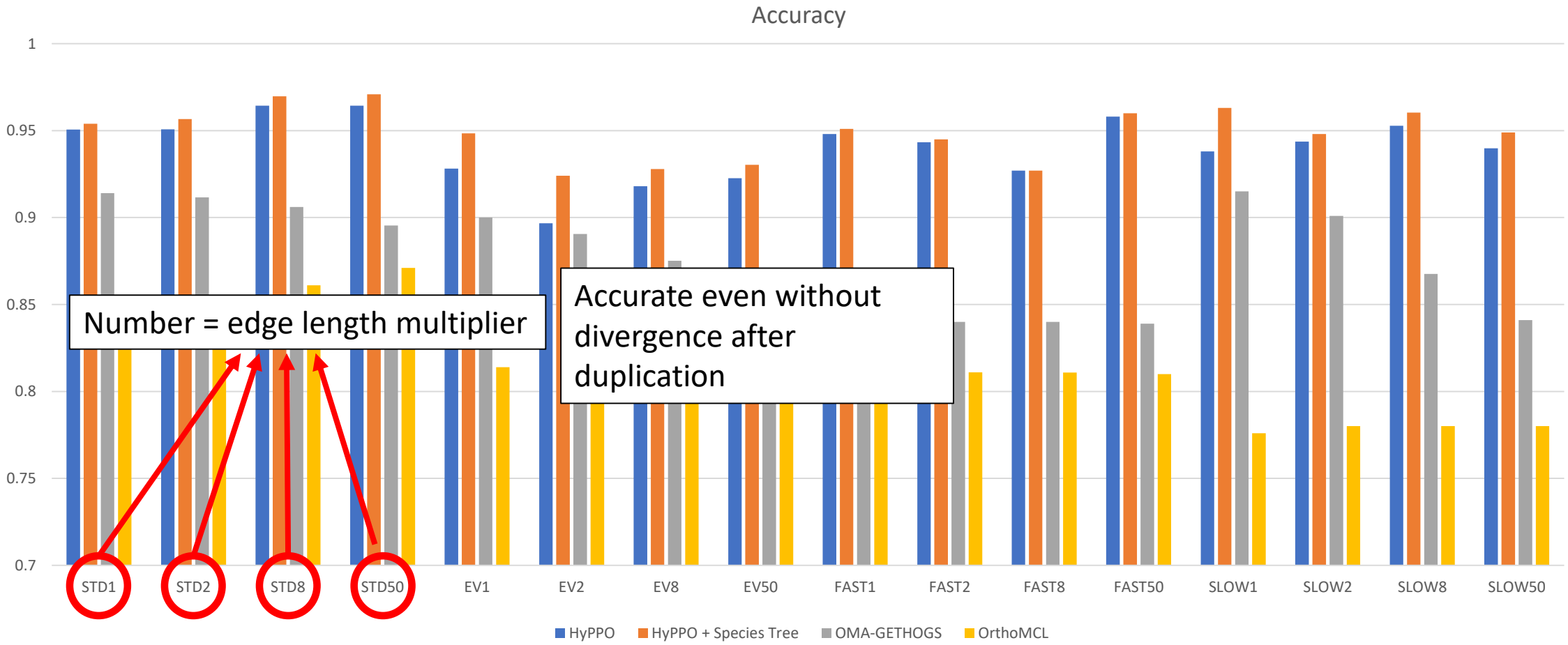
Accuracy on simulated trees

of correct relations / # of gene pairs



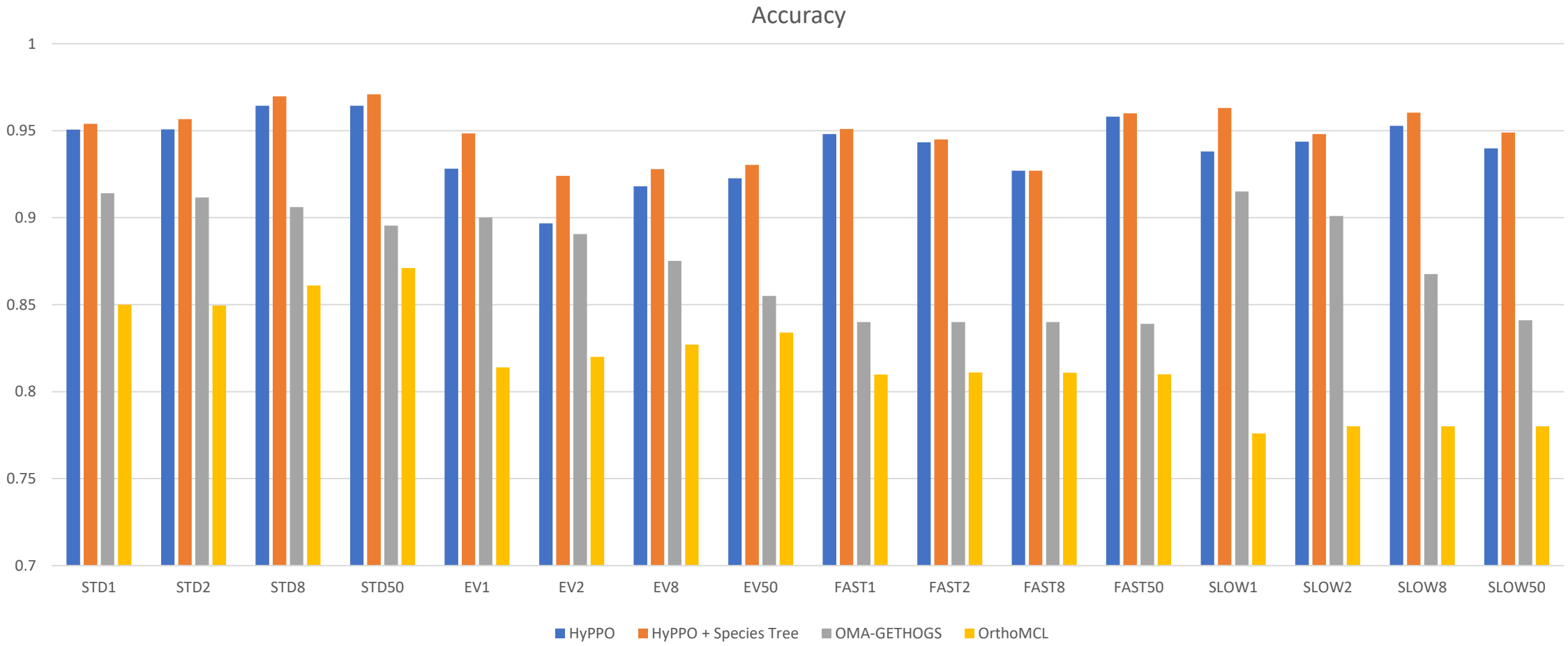
Accuracy on simulated trees

of correct relations / # of gene pairs



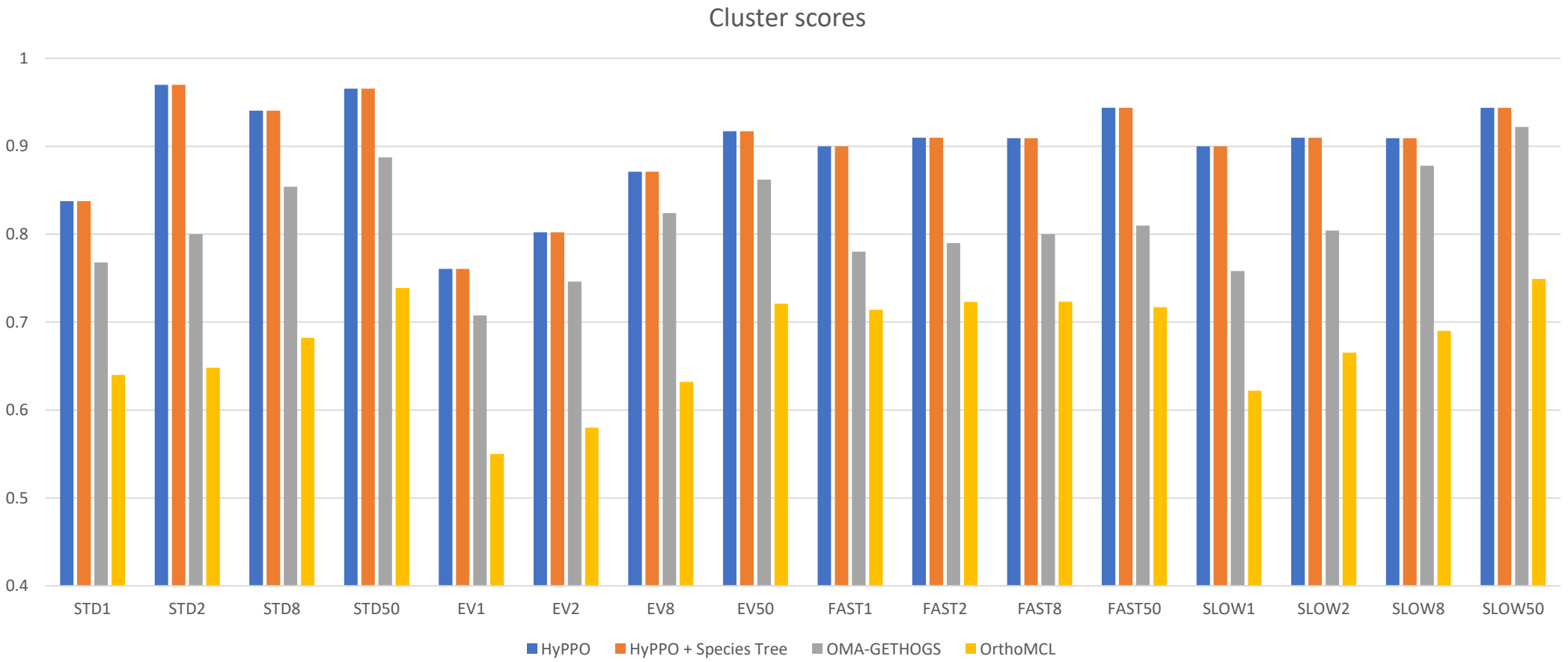
Accuracy on simulated trees

of correct relations / # of gene pairs



Quality of primary orthologs

cluster score = primary orthologs correctly together (see paper)



Averages on simulations

Precision = true pos / (true pos + false pos)

Recall = true pos / (true pos + false neg)

Method	Accuracy	Precision	Recall	Cluster score
HyPPO	.940	.911	.875	**.915**
HyPPO + species tree	**.949**	.924	**.905**	**.915**
OMA-GETHOGS	.877	**.940**	.699	.831
OrthoMCL	.812	.845	.496	.690

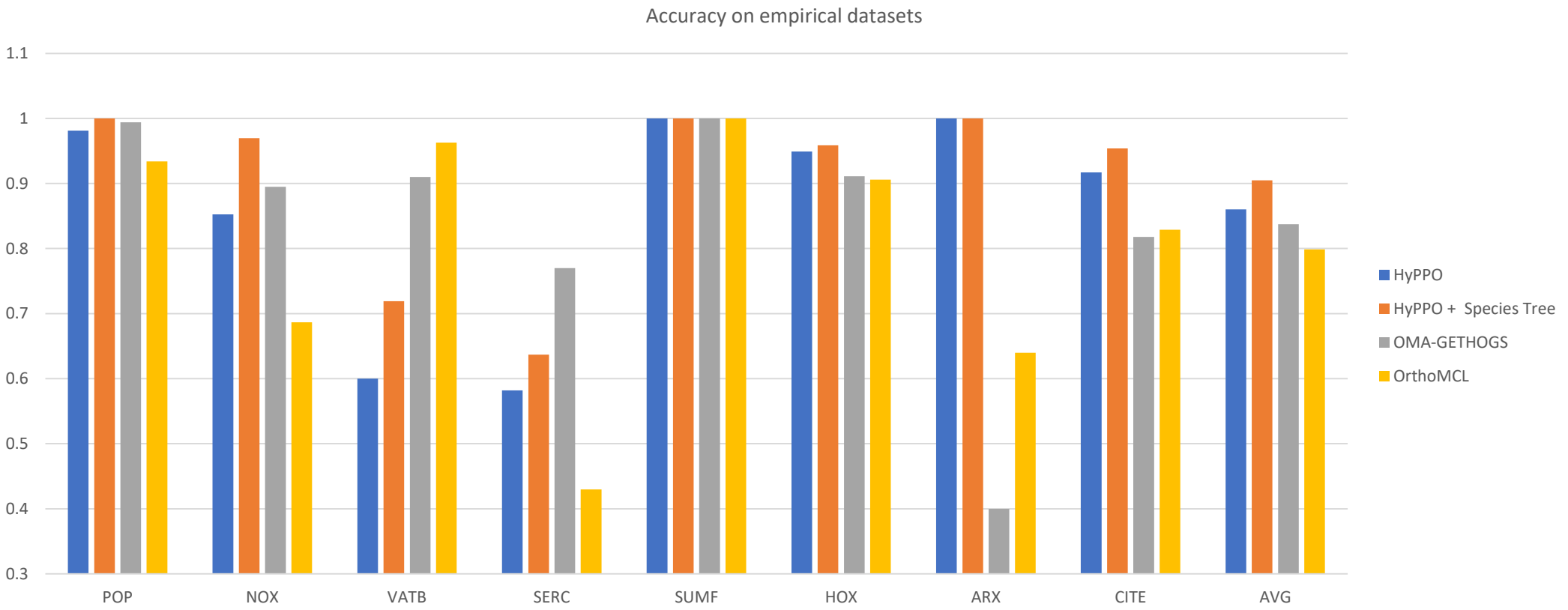
OMA has slightly better precision => less false pos
HyPPO has better recall => more orthologs found

SwissTree dataset

Gold standard manually curated gene trees

- 8 Eukaryote gene trees evaluated
  - POP, NOX, VATB, SERC, SUMF, HOX, ARX, CITE
- Speciation + duplication events known => orthologs/paralogs
- Primary orthologs not known

Accuracy on empirical datasets



HyPPO is bad on VATB because it has mostly orthologs (95% pairs), and HyPPO infers a duplication at the root and predicts ~50% paralogs.

Averages on SwissTrees

Precision = true pos / (true pos + false pos)

Recall = true pos / (true pos + false neg)

Method	Accuracy	Precision	Recall
HyPPO	.860	.941	.785
HyPPO + species tree	**.905**	.945	**.881**
OMA-GETHOGS	.837	**.950**	.711
OrthoMCL	.800	.859	.680

OMA has slightly better precision => less false pos
HyPPO has better recall => more orthologs found

Future work

Consider other post-duplication behavior.

- Duplication does not ***always*** introduce divergence.
- A more fine-grained classification of orthologs.

Make HyPPO more scalable.

Open algorithmic problem: reconstruct a species from orthology clusters.

- Main bottleneck in HyPPO.