The path of trial and tribulation towards homology based gene/species tree inference

N. Wieseke

Swarm Intelligence and Complex Systems Group

Leipzig University

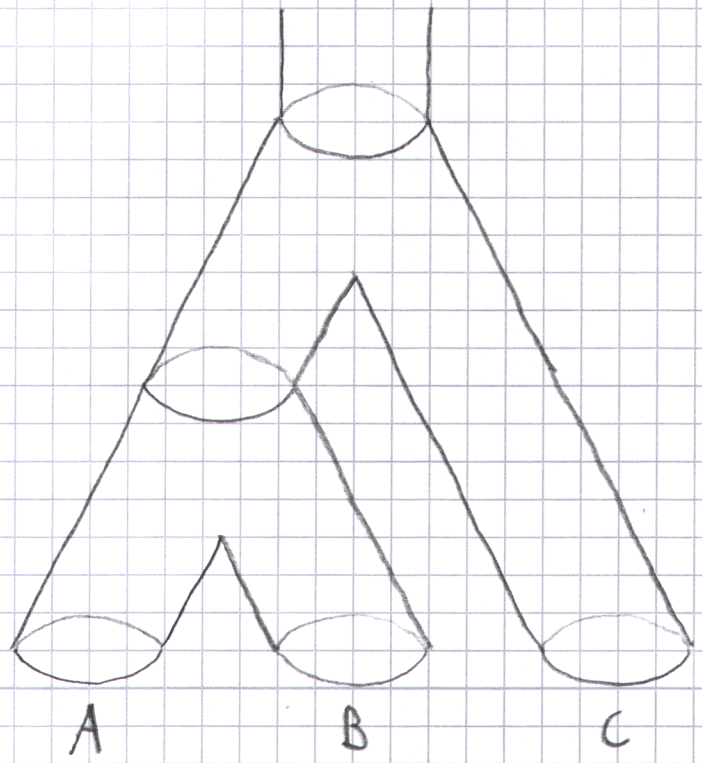
joint work with

Martin Middendorf, SICS, Leipzig

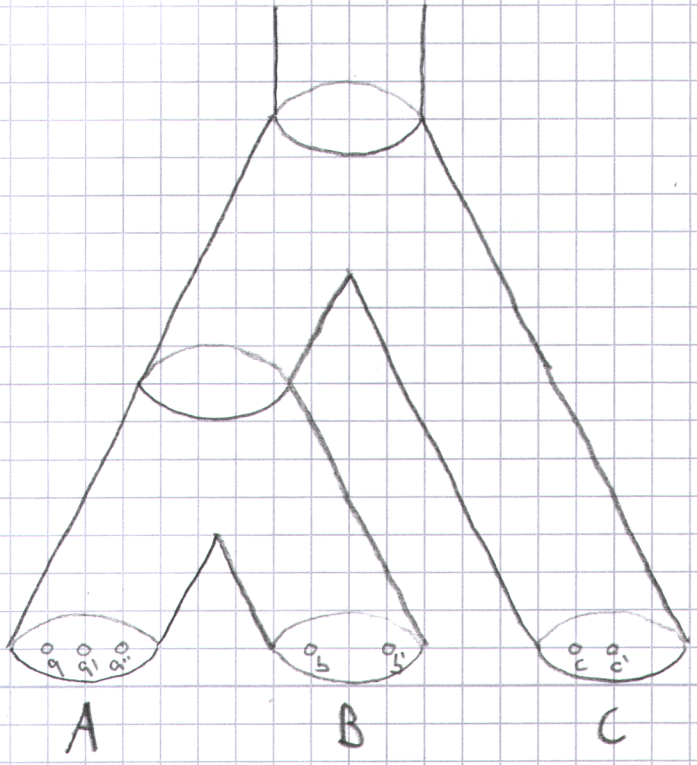
Peter F. Stadler, Bioinf, Leipzig

Marc Hellmuth, Bioinf, Greifswald

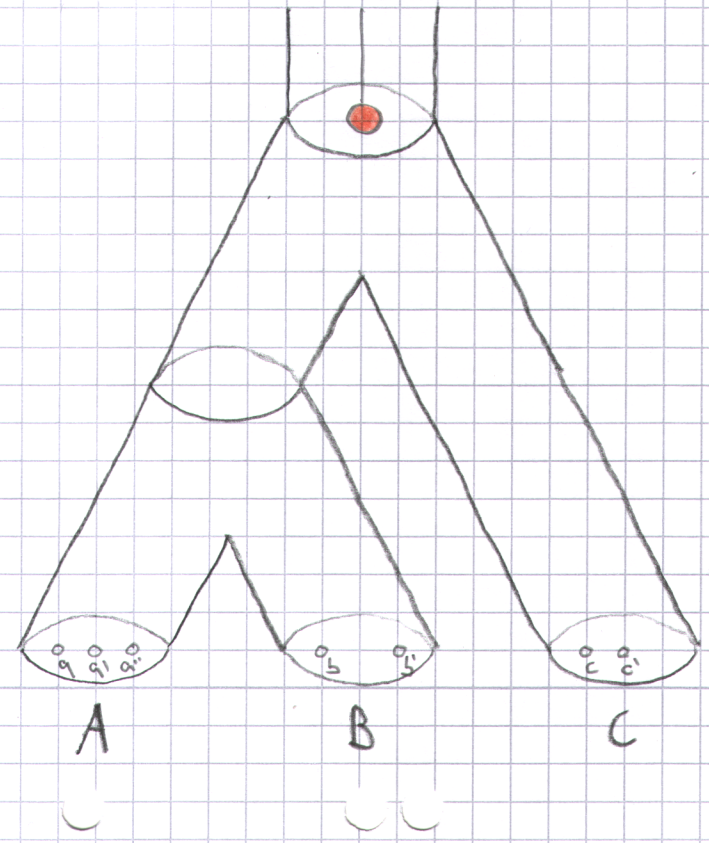
Problems in Phylogenetics



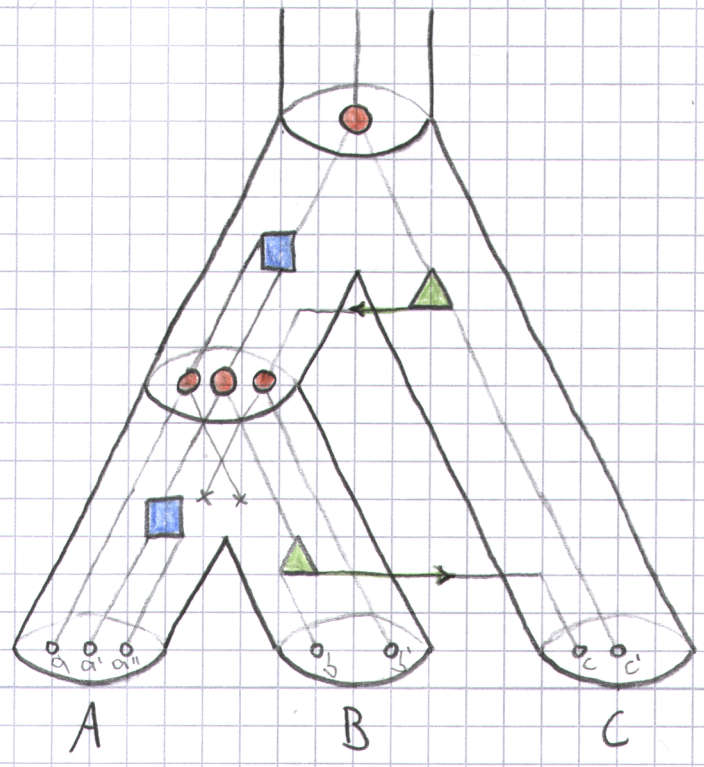
Problems in Phylogenetics



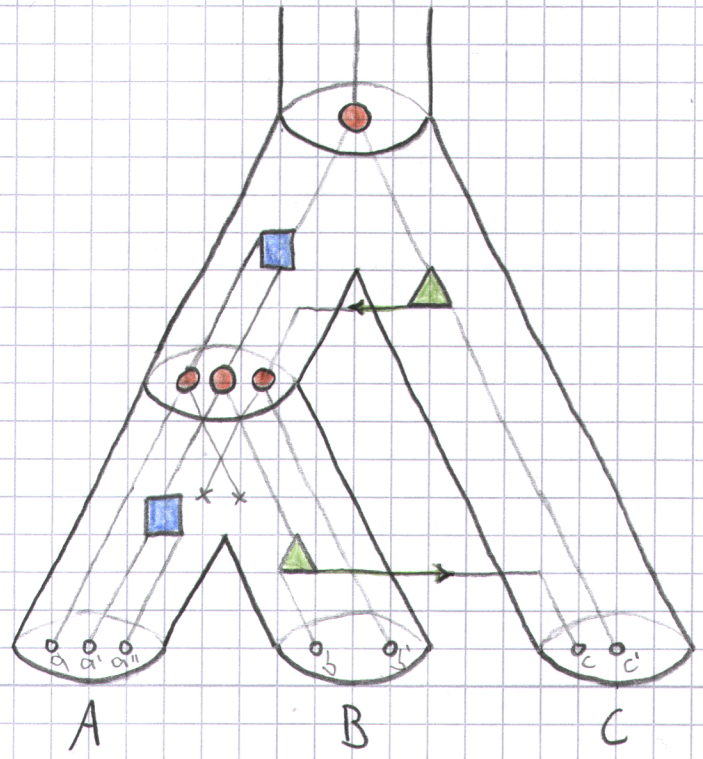
Problems in Phylogenetics



Problems in Phylogenetics



Problems in Phylogenetics



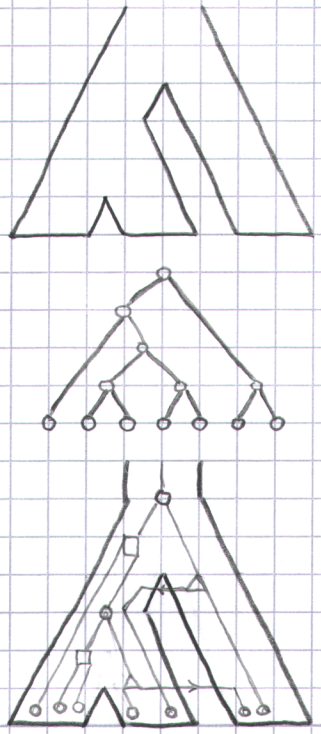
Phylogenetics asks for

- species tree
- gene trees
- evolutionary events
- reconciliation

using only the sequences from extant genes

Classical approach

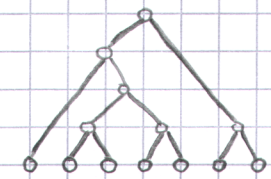
1. infer species tree
   - using 1-to-1 orthologs
   - multiple sequence alignments
   - NJ, ML, or Bayesian methods
2. infer gene tree
3. reconcile gene tree with species tree
   - using event costs or probabilities
4. obtain event labels



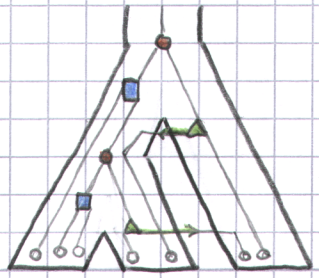
Classical approach

1. infer species tree
   - using 1-to-1 orthologs
   - multiple sequence alignments
   - NJ, ML, or Bayesian methods

2. infer gene tree



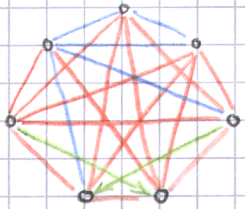
3. reconcile gene tree with species tree
   - using event costs or probabilities



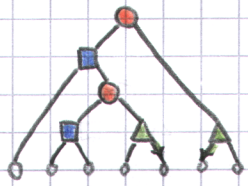
4. obtain event labels

New approach

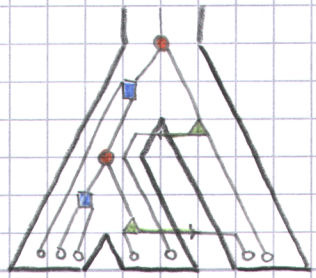
1. infer event labels
   - using pairwise sequence comparisons



2. obtain event labeled gene trees



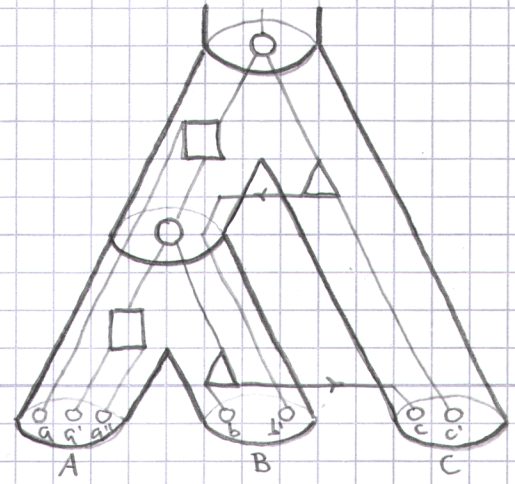
3. infer species tree and corresponding reconciliation



Homology Relations

LCA Relations

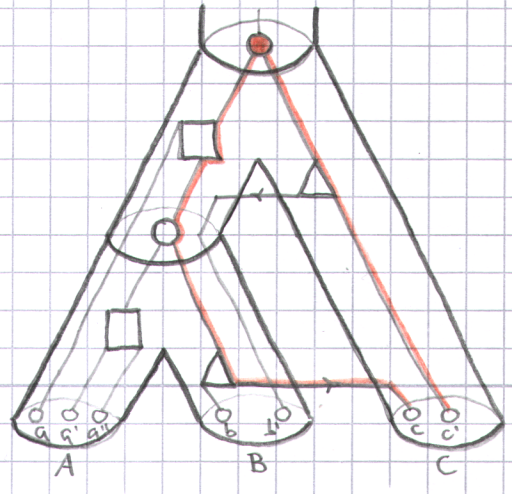
- lca-orthologs Θ_L
- lca-paralogs Π_L
- lca-xenologs χ_L
  - lca-donor-xenologs χ_L^D
  - lca-acceptor-xenologs χ_L^A



Homology Relations

LCA Relations

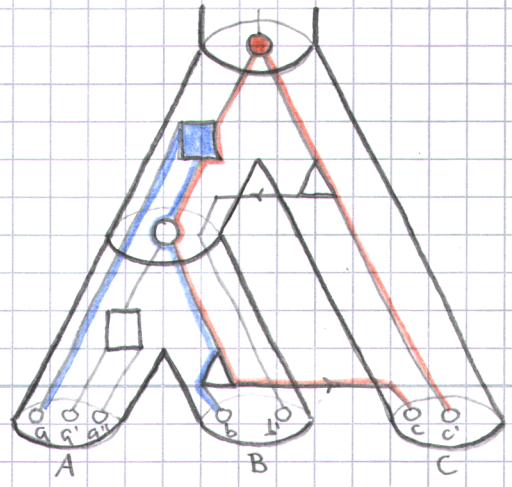
- lca - orthologs Θ_L
- lca - paralogs Π_L
- lca - xenologs χ_L
  - lca - donor - xenologs χ_L^D
  - lca - acceptor - xenologs χ_L^A



Homology Relations

LCA Relations

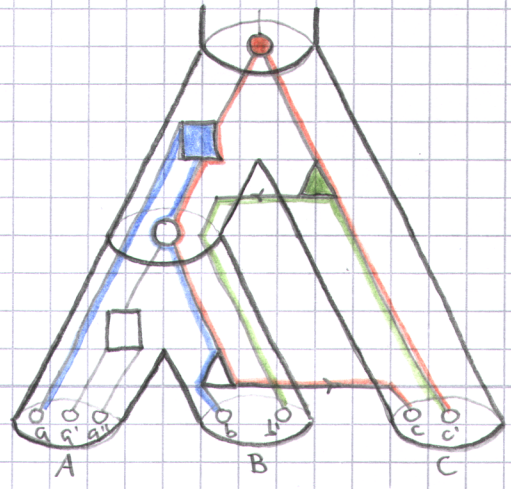
- lca-orthologs Θ_L
- lca-paralogs Π_L
- lca-xenologs χ_L
  - lca-donor-xenologs χ_L^D
  - lca-acceptor-xenologs χ_L^A



Homology Relations

LCA Relations

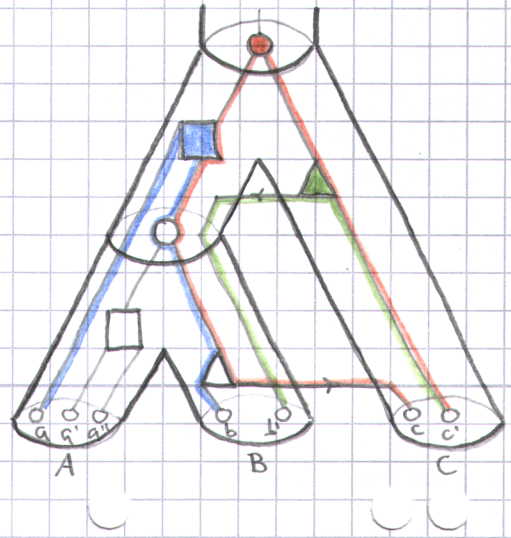
- lca - orthologs Θ_L
- lca - paralogs Π_L
- lca - xenologs χ_L
  - lca - donor - xenologs χ_L^D
  - lca - acceptor - xenologs χ_L^A



Homology Relations

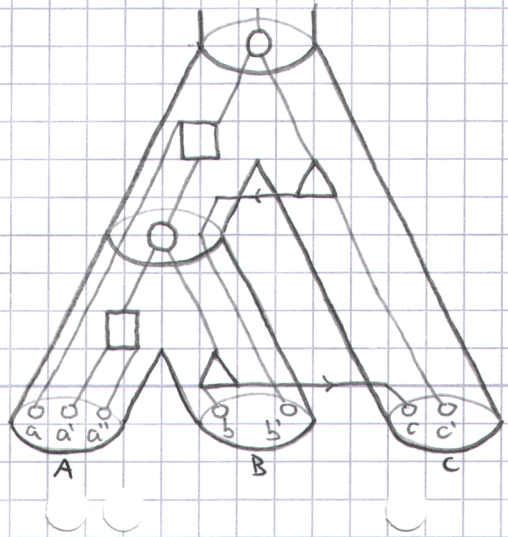
LCA Relations

- lca-orthologs Θ_L
- lca-paralogs Π_L
- lca-xenologs χ_L
  - lca-donor-xenologs χ_L^D
  - lca-acceptor-xenologs χ_L^A



Fitch Relations

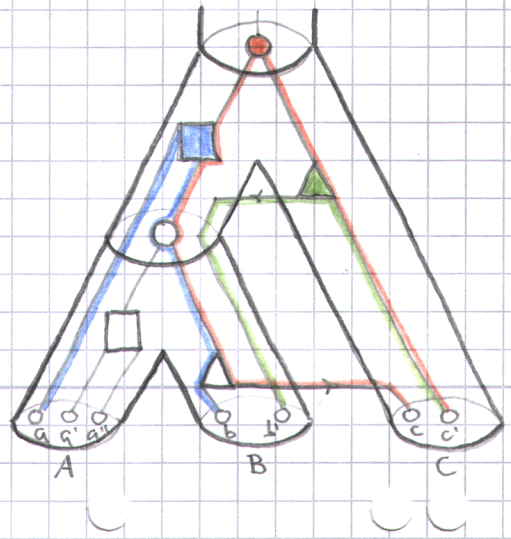
- Fitch-orthologs Θ_F
  $(a,b) \in \Theta_F \leftrightarrow \mu(lca_T(a,b)) = lca_S(A,B)$
- Fitch-xenologs χ_F
  $(a,b) \in \chi_F \leftrightarrow$ transfer on path $a-b$
  - Fitch-donor-xenologs χ_F^D
  - Fitch-acceptor-xenologs χ_F^A



Homology Relations

LCA Relations

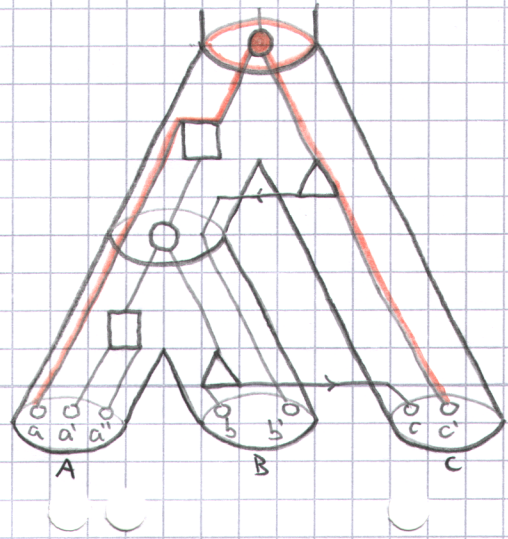
- lca-orthologs Θ_L
- lca-paralogs Π_L
- lca-xenologs χ_L
  - lca-donor-xenologs χ_L^D
  - lca-acceptor-xenologs χ_L^A



$\Theta_L \supseteq \Theta_F$

Fitch Relations

- Fitch-orthologs Θ_F
  $(a,b) \in \Theta_F \leftrightarrow \mu(lca_T(a,b)) = lca_S(A,B)$
- Fitch-xenologs χ_F
  $(a,b) \in \chi_F \leftrightarrow$ transfer on path a-b
  - Fitch-donor-xenologs χ_F^D
  - Fitch-acceptor-xenologs χ_F^A



Homology Relations

LCA Relations

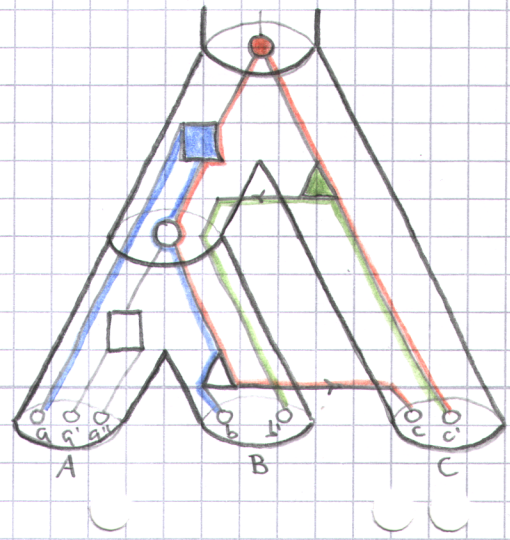
- lca-orthologs Θ_L
- lca-paralogs Π_L
- lca-xenologs χ_L
  - lca-donor-xenologs χ_L^D
  - lca-acceptor-xenologs χ_L^A

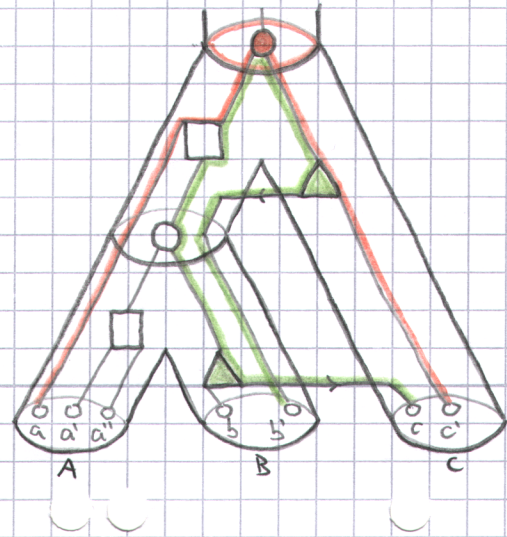
$\Theta_L \supseteq \Theta_F$

$\chi_L \subseteq \chi_F$

Fitch Relations

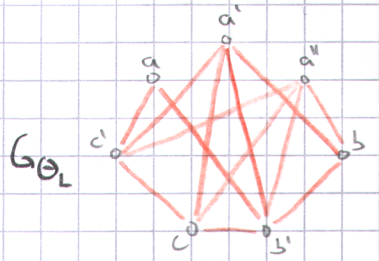
- Fitch-orthologs Θ_F
  $(a,b) \in \Theta_F \leftrightarrow M(lca_T(a,b)) = lca_S(A,B)$
- Fitch-xenologs χ_F
  $(a,b) \in \chi_F \leftrightarrow$ transfer on path a-b
  - Fitch-donor-xenologs χ_F^D
  - Fitch-acceptor-xenologs χ_F^A

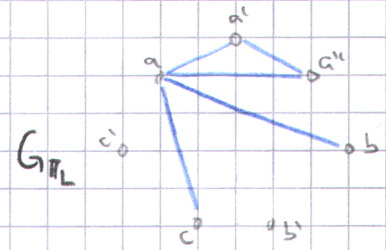


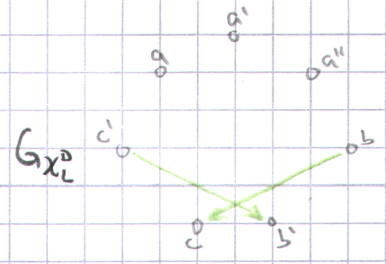


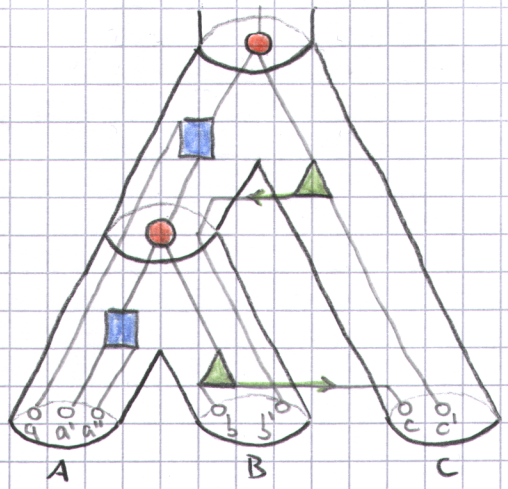
Graph Representation

LCA Relations

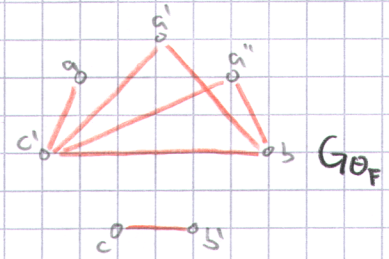


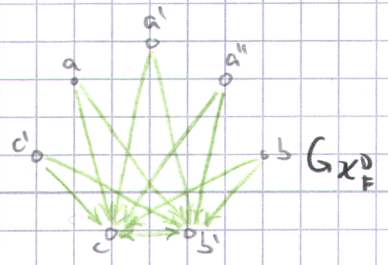






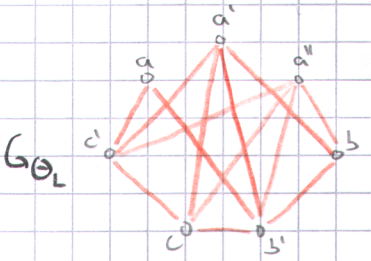
Fitch Relations

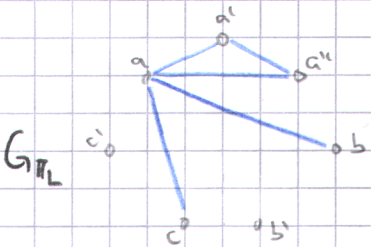


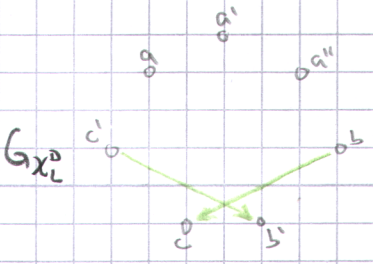


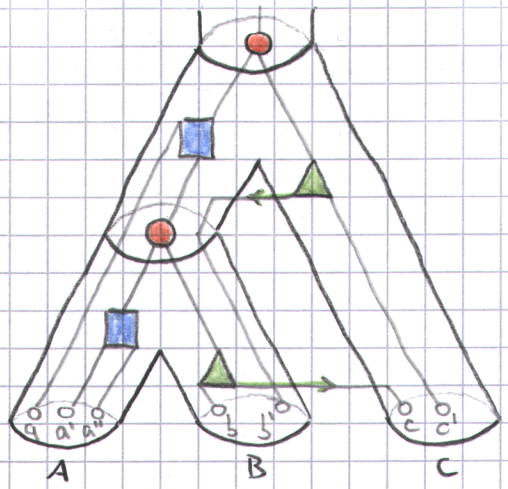
Graph Representation

LCA Relations







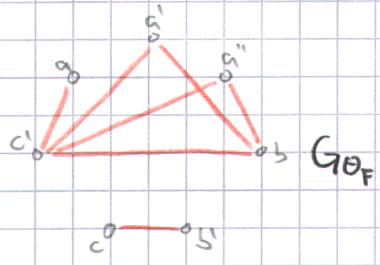


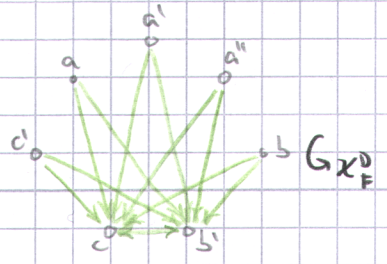
In case of no HGT events

$$G_{\Theta_L} \cong G_{\Theta_F}$$

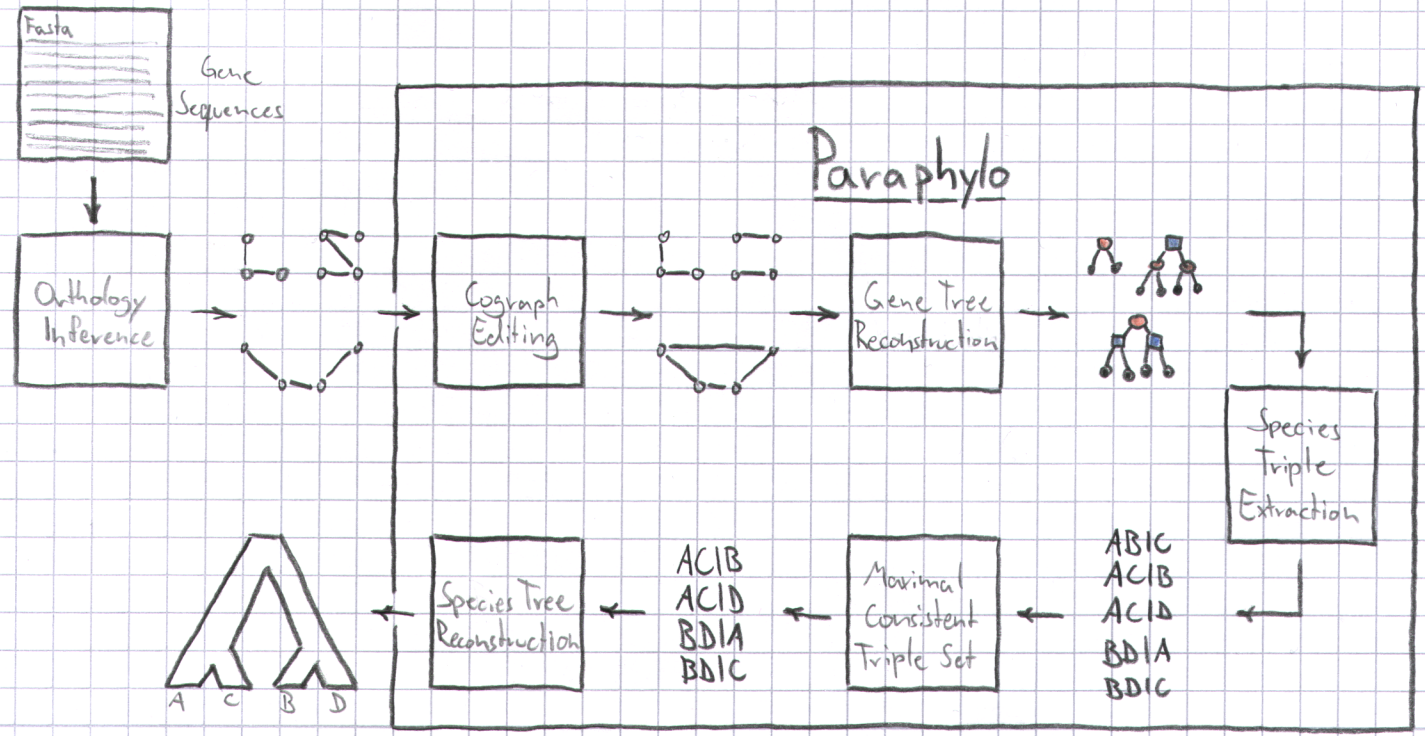
$$G_{\Pi_L} \cong \overline{G_{\Theta_F}}$$

Fitch Relations

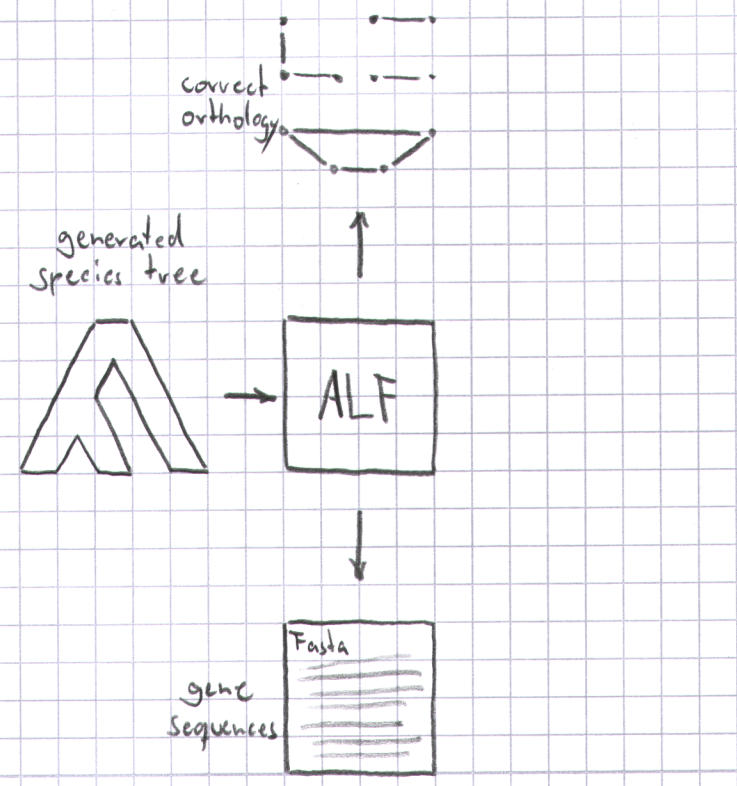




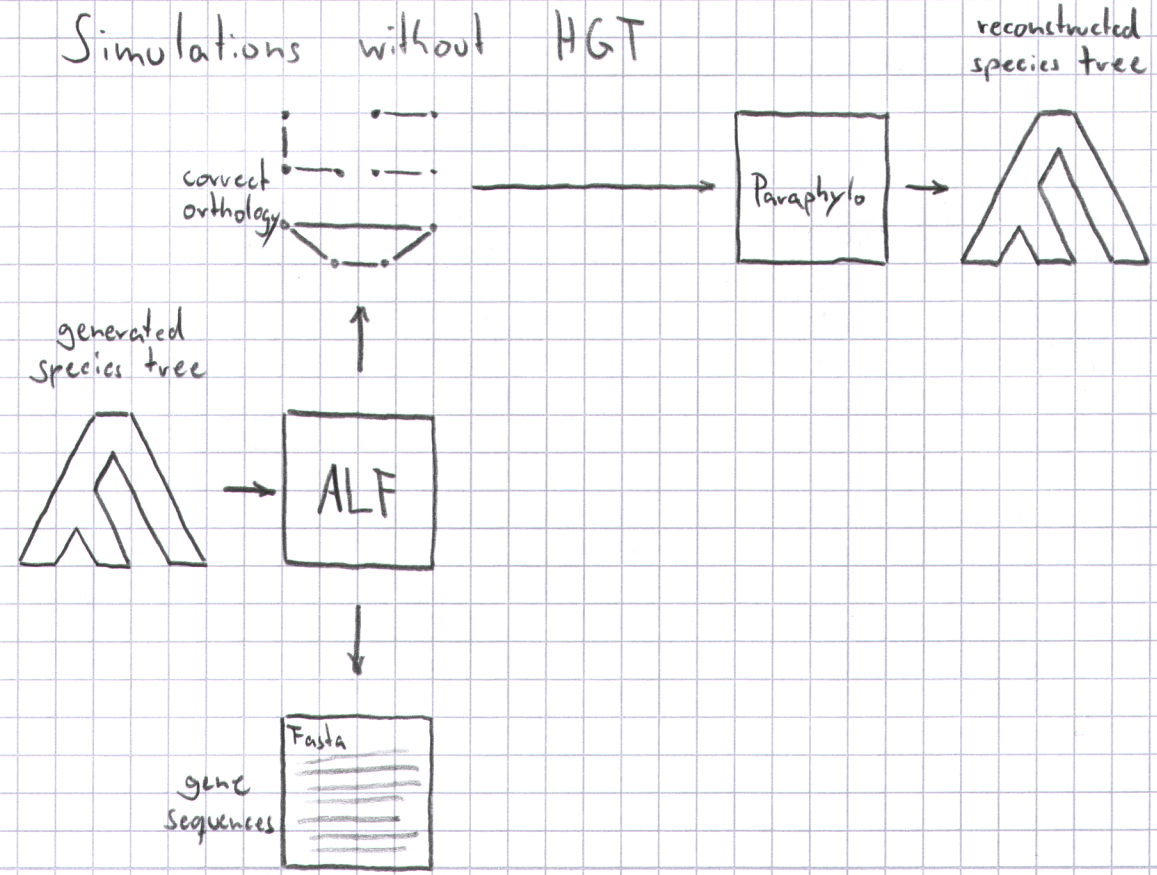
Pipeline for datasets without HGT



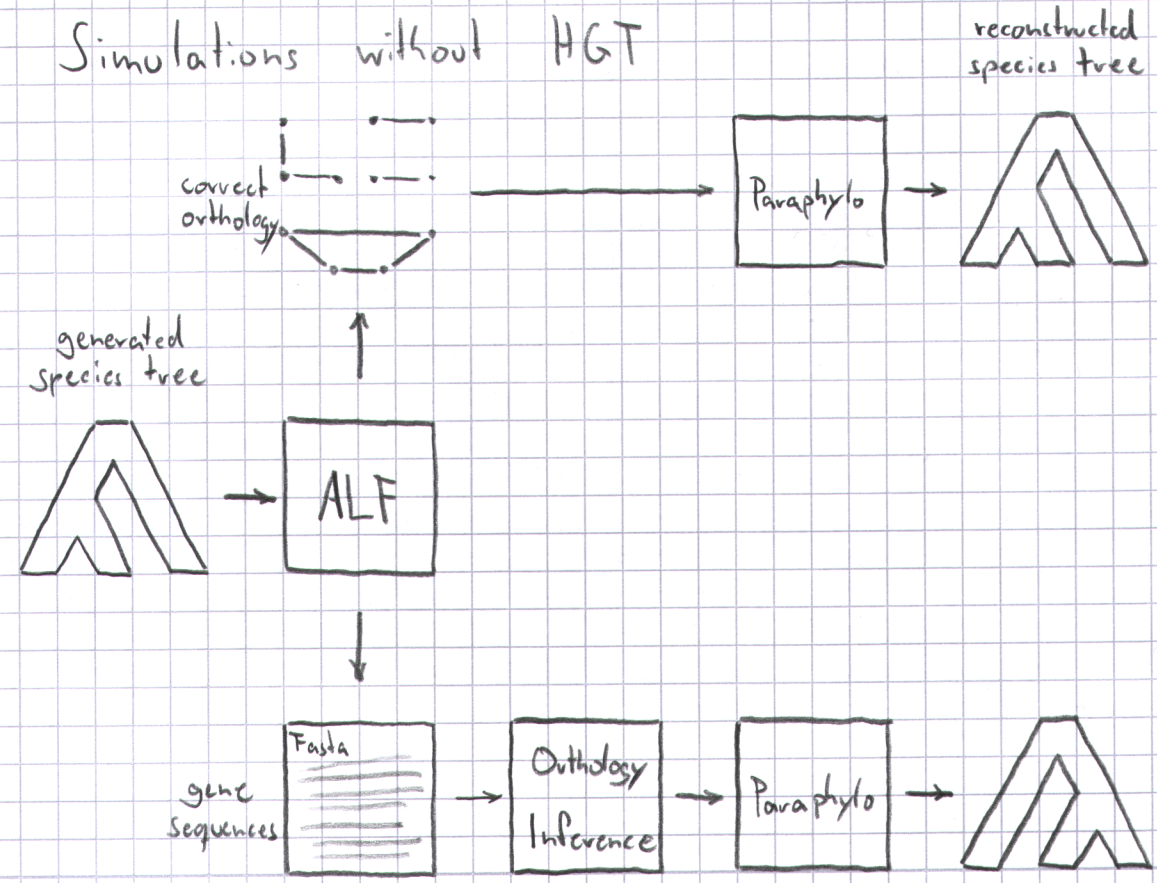
Simulations without HGT



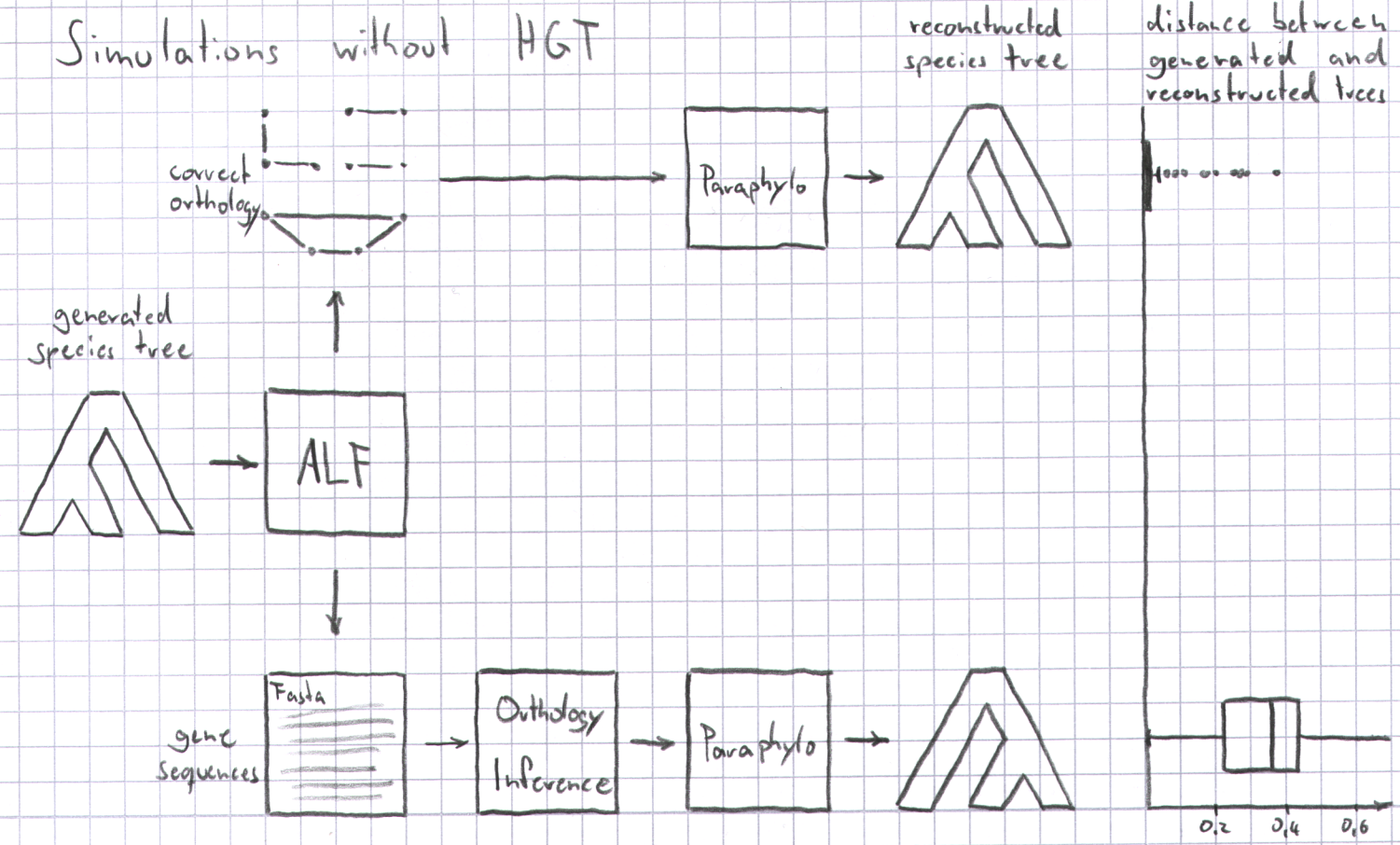
Simulations without HGT



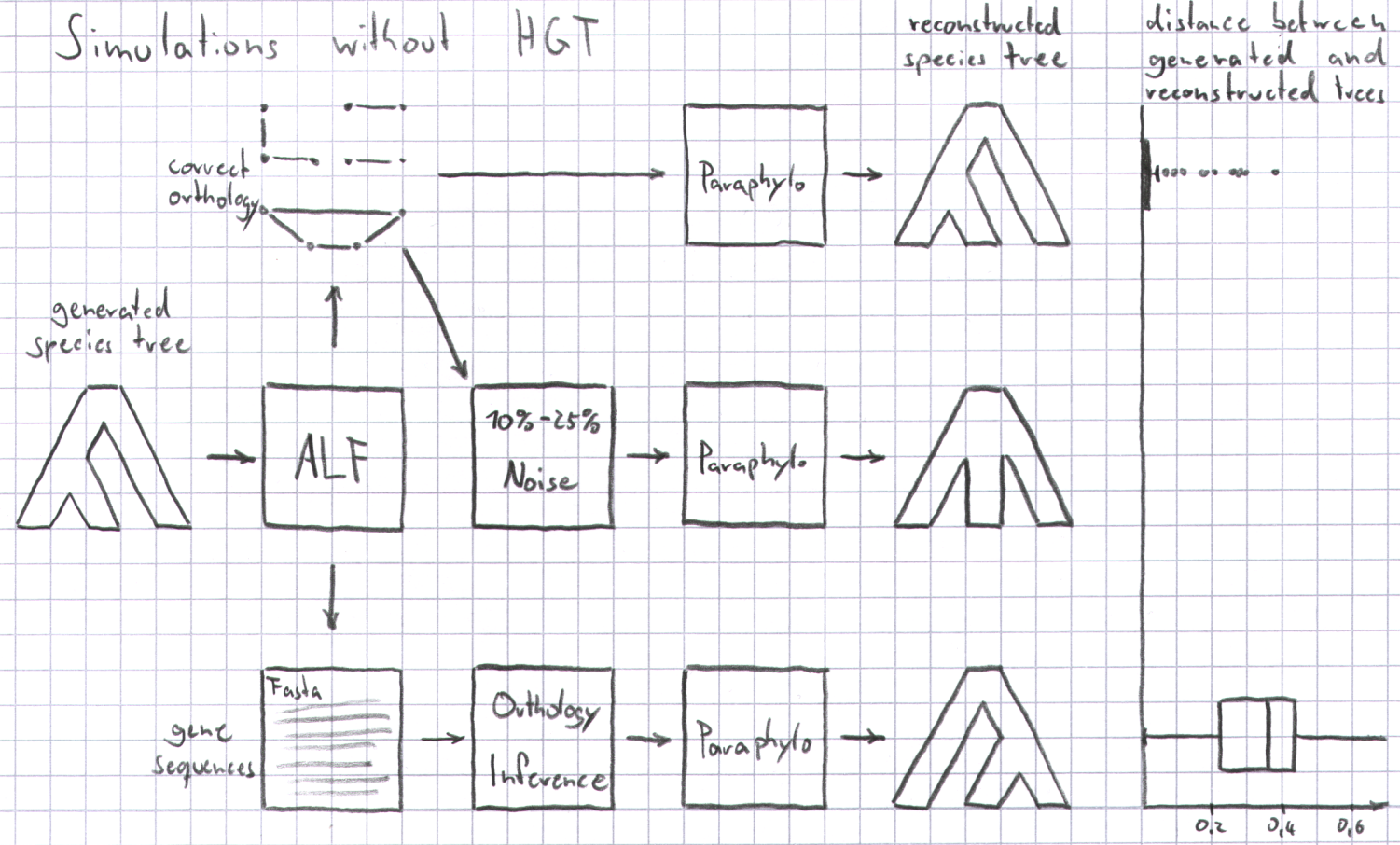
Simulations without HGT



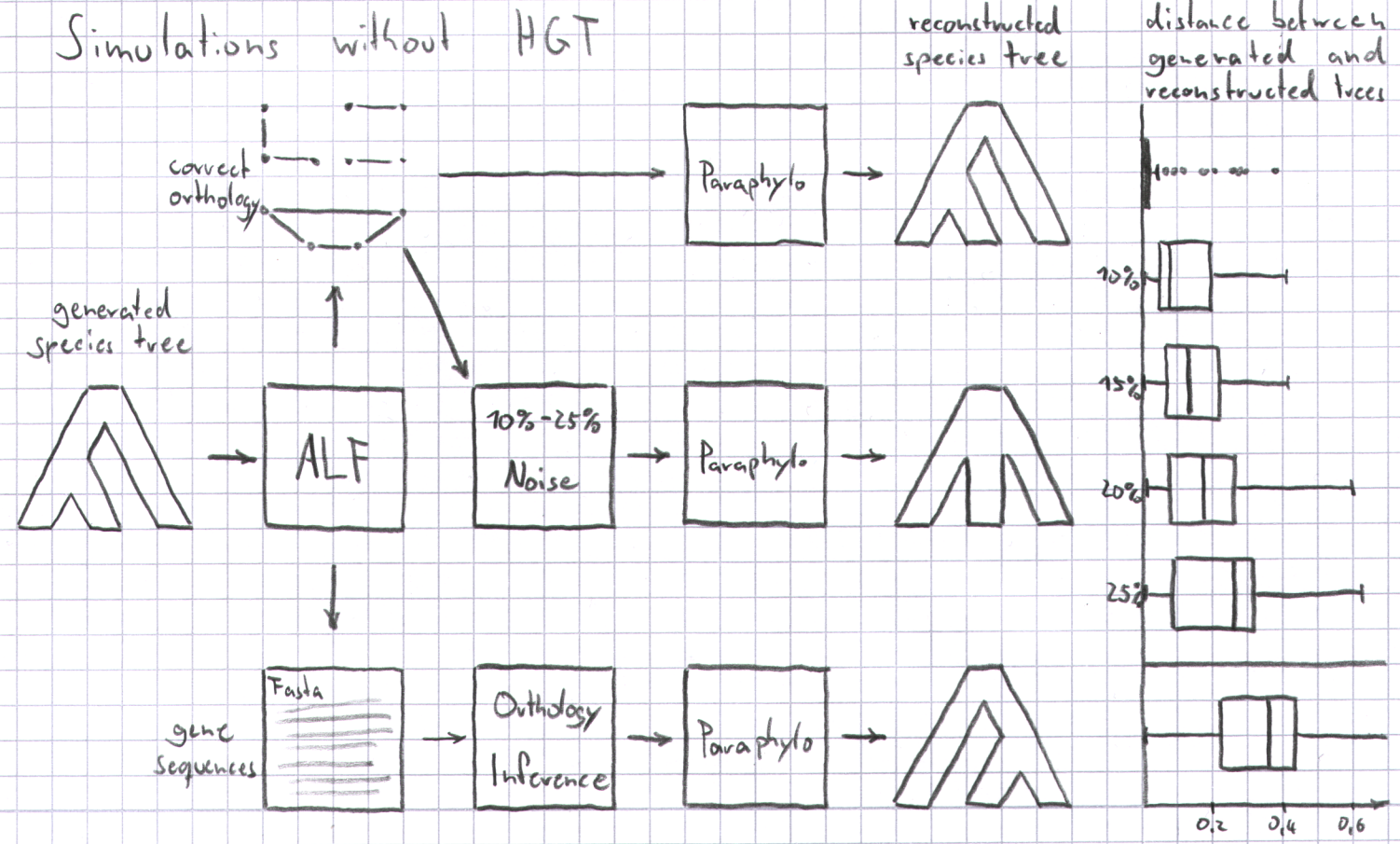
Simulations without HGT



Simulations without HGT



Simulations without HGT



Orthology Inference - when does it go wrong

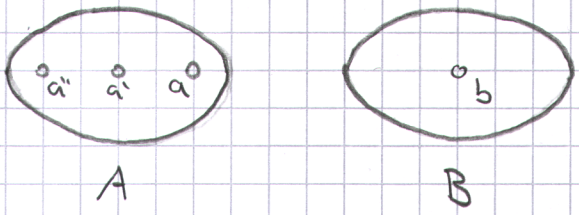
- ☐ no molecular clock-like data

Orthology Inference - when does it go wrong

- ☐ no molecular clock-like data
- ☐ horizontal gene transfer
- ☐ gene losses

Orthology Inference - based on pairwise sequence comparisons

- compute evolutionary distance $d(a,b)$ between all genes a, b
- determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- consider RBMs as orthologs

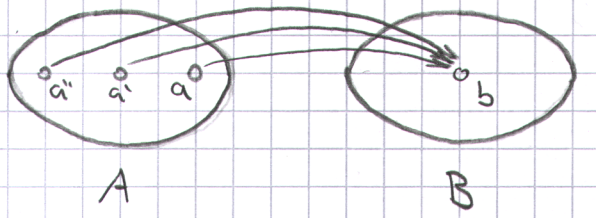


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches

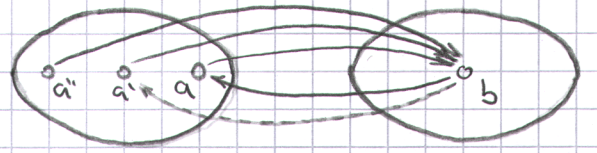
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$

  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs



Orthology Inference – based on pairwise sequence comparisons

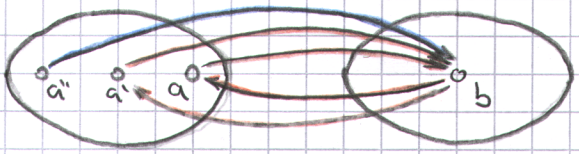
- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☑ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs



$d(a,b) \approx d(a',b) \leq d(a'',b)$

Orthology Inference – based on pairwise sequence comparisons

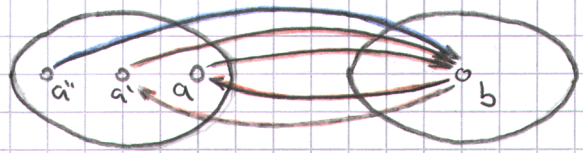
- compute evolutionary distance $d(a,b)$ between all genes a, b
- determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- consider RBMs as orthologs



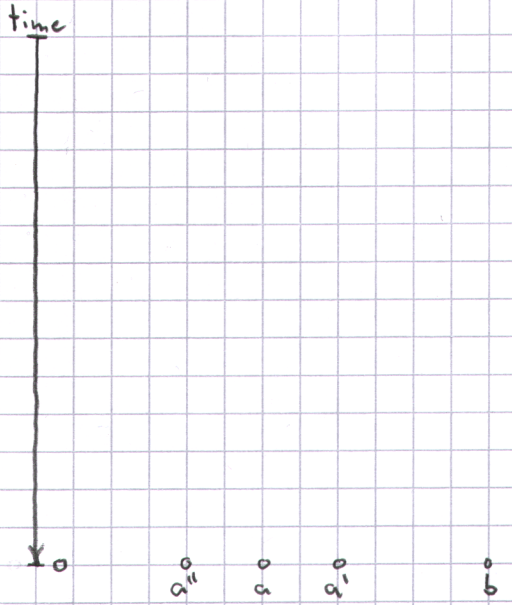
$d(a,b) \approx d(a',b) \ll d(a'',b)$

Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs

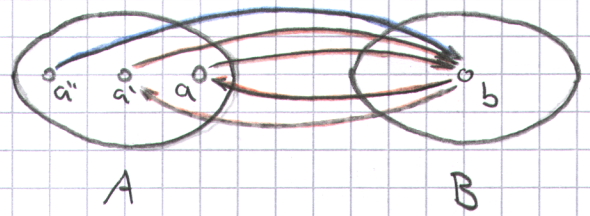


$d(a,b) \approx d(a',b) \ll d(a'',b)$

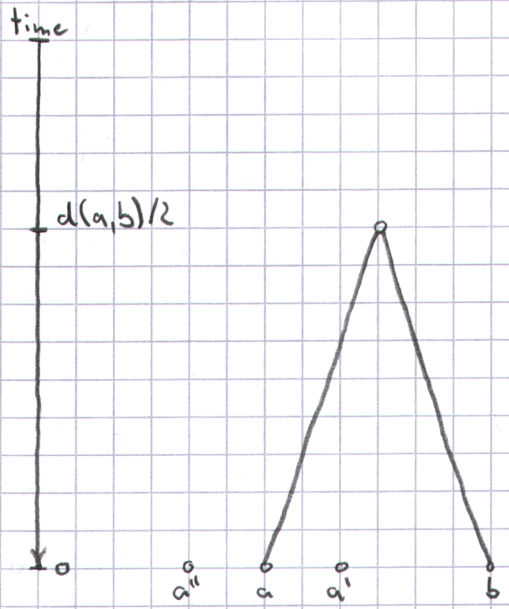


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs

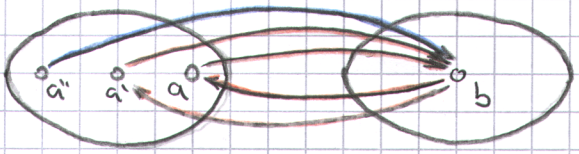


$d(a,b) \approx d(a',b) \ll d(a'',b)$

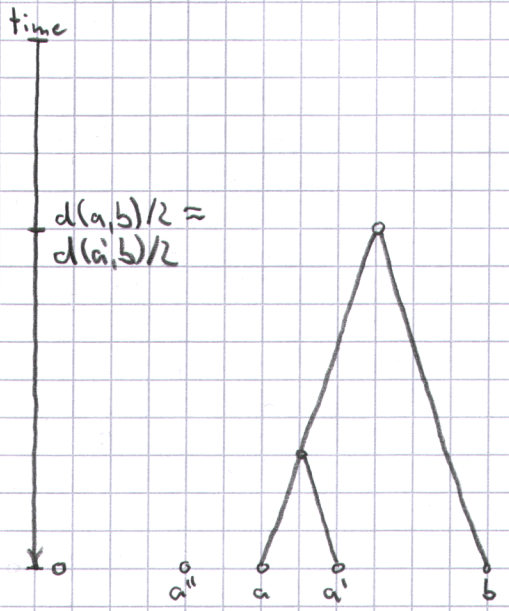


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs

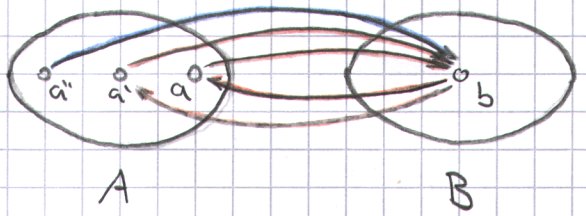


$d(a,b) \approx d(a',b) \lessapprox d(a'',b)$

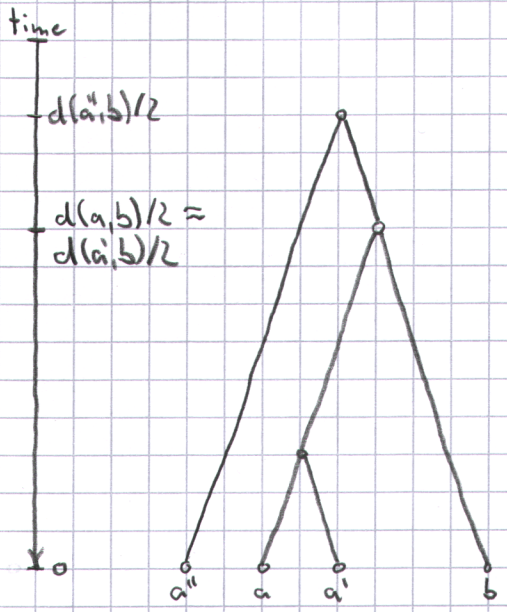


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs

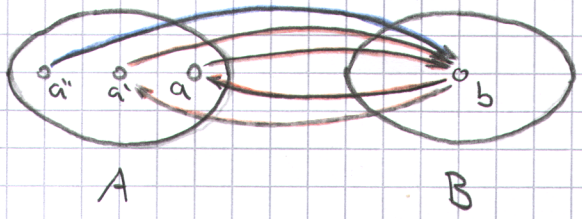


$d(a,b) \approx d(a',b) \ll d(a'',b)$

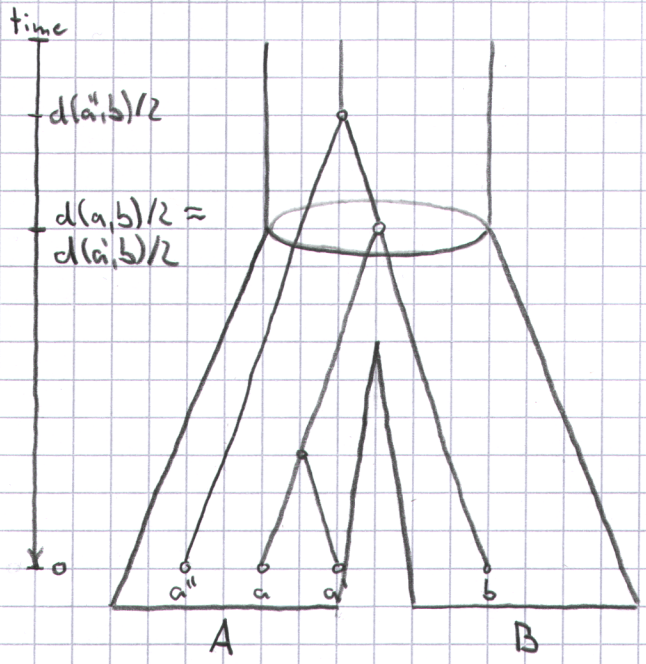


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs

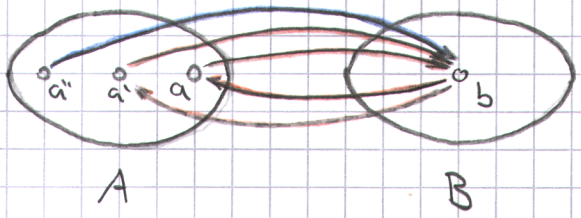


$d(a,b) \approx d(a',b) \leq d(a'',b)$

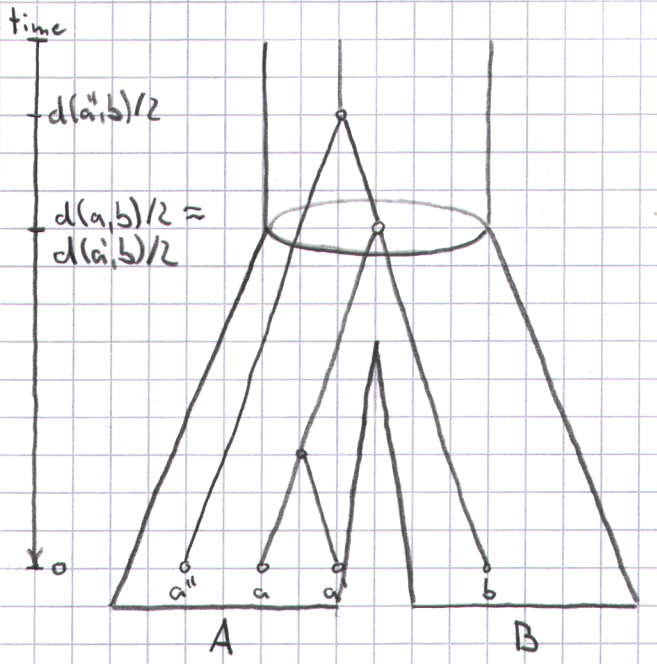


Orthology Inference – based on pairwise sequence comparisons

- ☐ compute evolutionary distance $d(a,b)$ between all genes a, b
- ☐ determine reciprocal best matches
  (a,b) is a RBM iff $\forall a' \in A, b' \in B$
  $d(a,b) \leq d(a',b)$ and $d(a,b) \leq d(a,b')$
- ☐ consider RBMs as orthologs



$d(a,b) \approx d(a',b) \ll d(a'',b)$



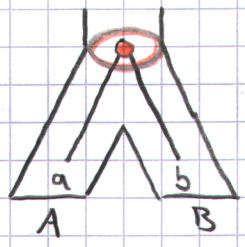
If (a,b) is a RBM then assume

$d(a,b) \triangleq d(A,B)$

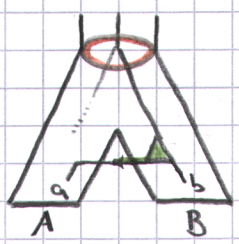
Orthology Inference - when does it go wrong

Let (a,b) be a RBM

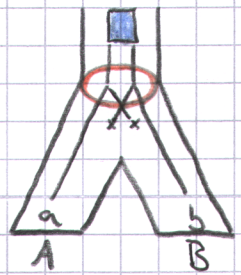
- $d(a,b) \hat{=} d(A,B)$



- $d(a,b) \hat{<} d(A,B)$



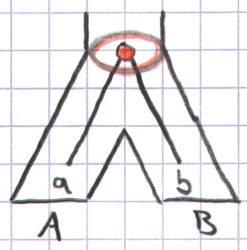
- $d(a,b) \hat{>} d(A,B)$



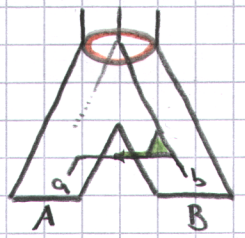
Orthology Inference - when does it go wrong

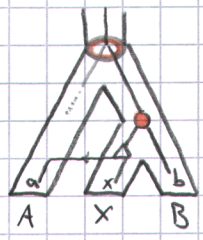
Let (a,b) be a RBM

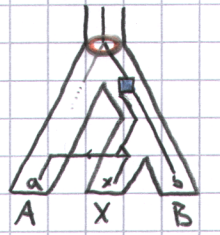
- $d(a,b) \hat{=} d(A,B)$



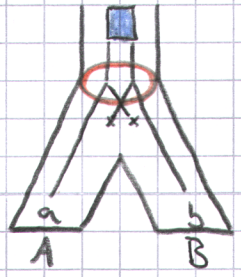
- $d(a,b) \hat{<} d(A,B)$

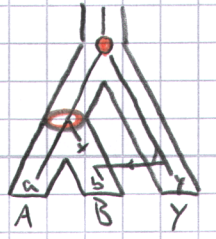


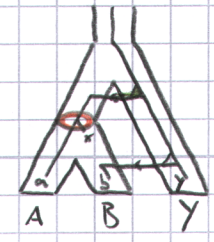




- $d(a,b) \hat{>} d(A,B)$

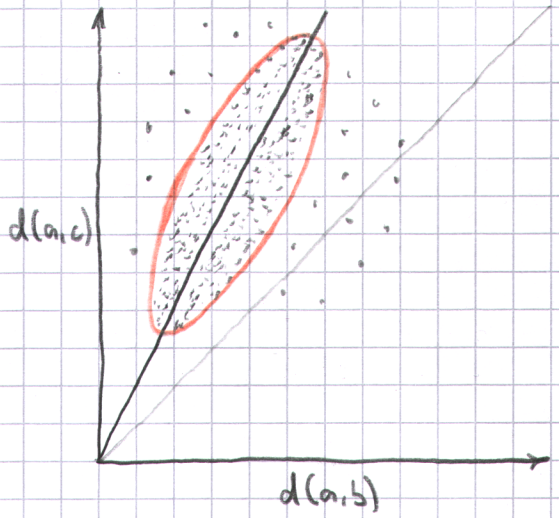


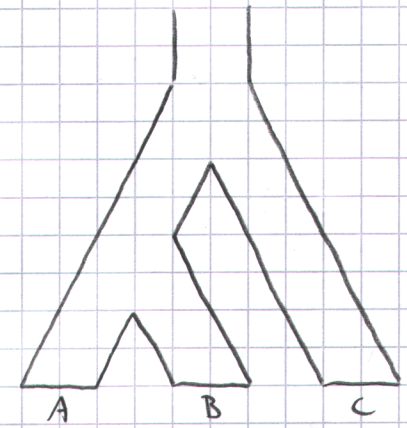




Improved Orthology Inference

- use RBM strategy to obtain an initial orthology estimation

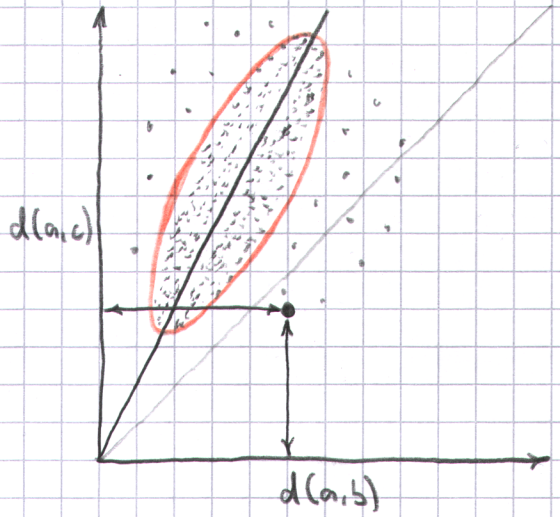


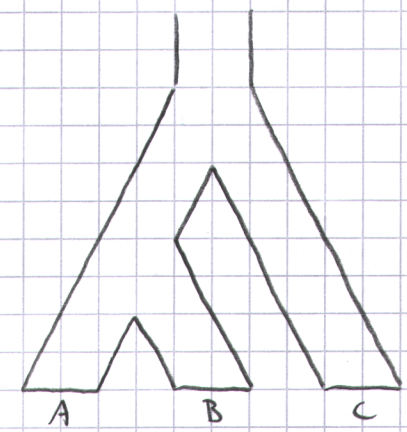


- remove (a,b), (a,c) for all outliers $(d(a,b), d(a,c))$

Improved Orthology Inference

- use RBM strategy to obtain an initial orthology estimation

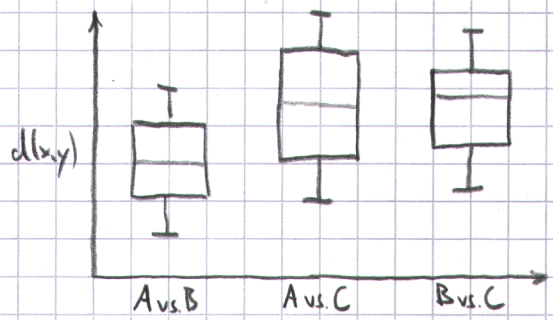


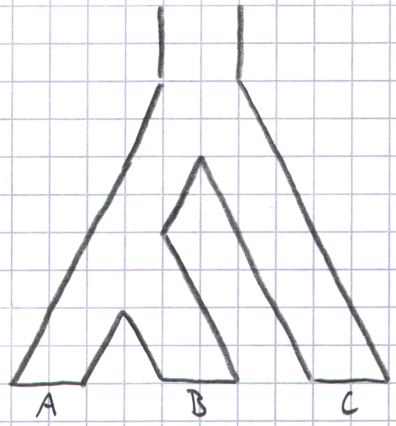


- remove (a,b), (a,c) for all outliers $(d(a,b), d(a,c))$

Improved Orthology Inference

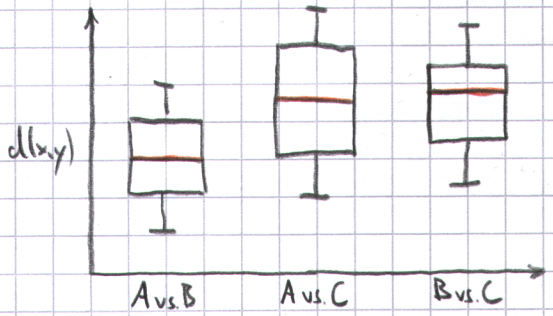
- use filtered orthology estimates



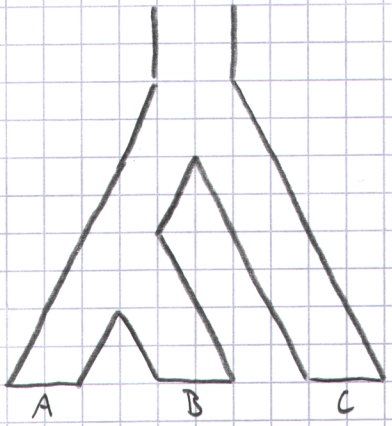


Improved Orthology Inference

- use filtered orthology estimates

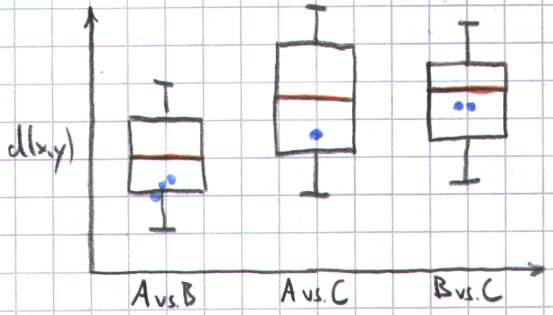


- define $d(X,Y) := \text{median}\left(d(x,y)\right)$

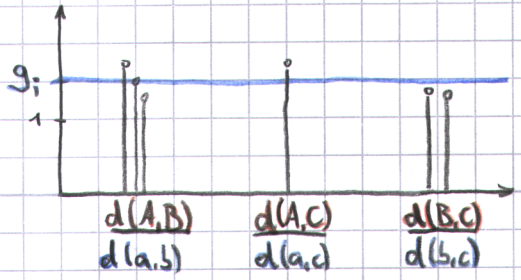


Improved Orthology Inference

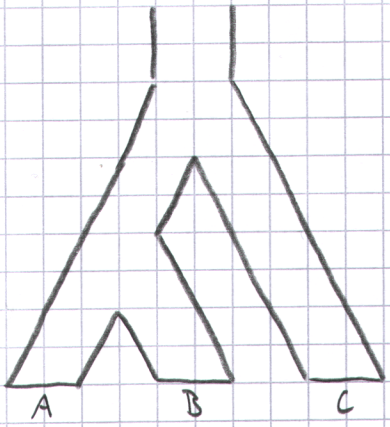
- use filtered orthology estimates



- define $d(X,Y) := \text{median}\,(d(x,y))$



with a,b,c from gene family i



Improved Orthology Inference

- $d(a,b) \cong d(A,B)$ translates to $d(a,b) \cdot g_i \approx d(A,B)$
- $d(a,b) \hat{<} d(A,B)$ — " — $d(a,b) \cdot g_i \ll d(A,B)$
- $d(a,b) \hat{>} d(A,B)$ — " — $d(a,b) \cdot g_i \gg d(A,B)$

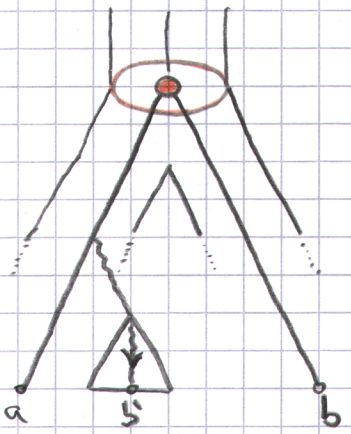
Infer Fitch Orthologs Θ_F

$$\Theta_F = \{(a,b) \mid d(a,b) \cdot g_i \approx d(A,B)\}$$

Fitch Xenology Inference

Idea: $(a,b') \in \chi^D_F$ if $\exists\, b$ s.t. $d(a,b') \cdot g_i \ll d(A,B) \approx d(a,b) \cdot g_i$

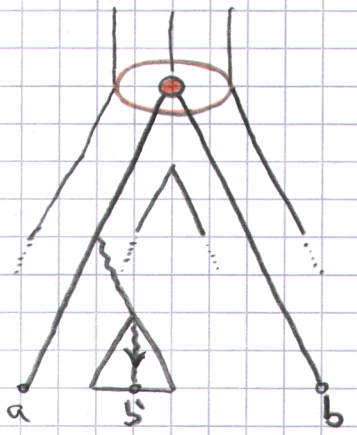
$$\chi^D_F = \{(x,b') \mid \exists\, a,b,\ a \in A, b,b' \in B,\ d(x,a) \lesssim d(x,b') \text{ and } d(a,b') \cdot g_i \ll d(A,B) \approx d(a,b) \cdot g_i\}$$



Fitch Xenology Inference

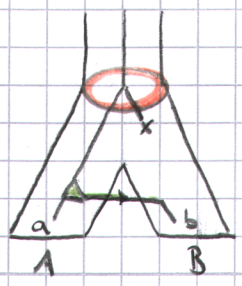
Idea: $(a,b') \in \chi^{D}_{F}$ if $\exists\, b$ s.t. $d(a,b') \cdot g_i \ll d(A,B) \approx d(a,b) \cdot g_i$

$\chi^{D}_{F} = \{(x,b') \mid \exists\, a,b,\ a \in A, b,b' \in B,\ d(x,a) \lesssim d(x,b')$ and $d(a,b') \cdot g_i \ll d(A,B) \approx d(a,b) \cdot g_i\}$

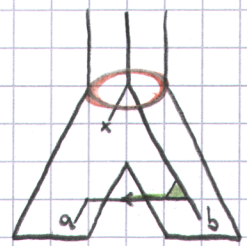


There is ambiguity!

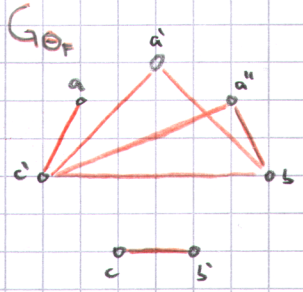
$(a,b) \in \chi^{D}_{F}$

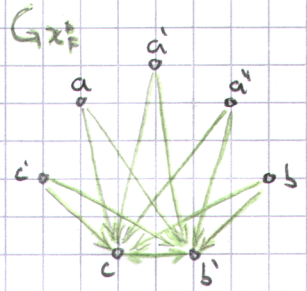


$(b,a) \in \chi^{D}_{F}$

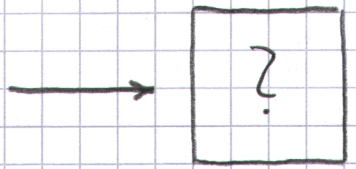


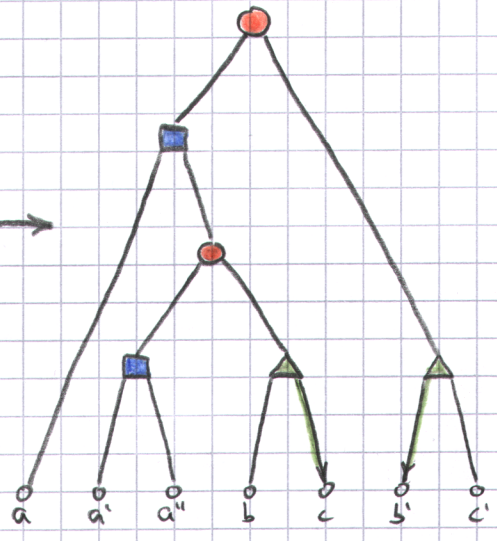
Reconstruction of event labeled gene trees





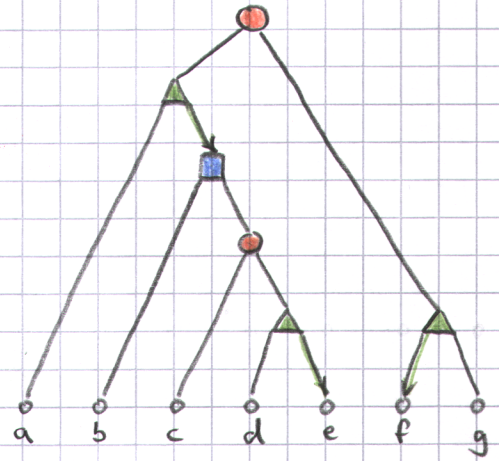
estimated Fitch Relations

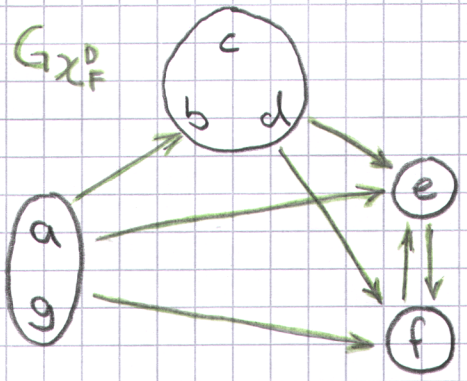




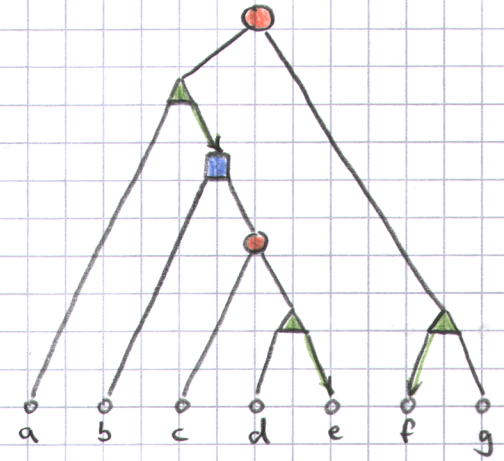
event labeled gene tree

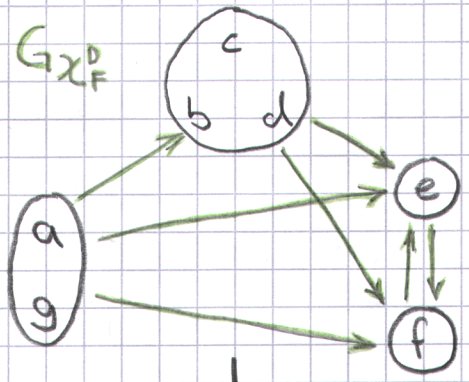
Reconstruction of event labeled gene trees



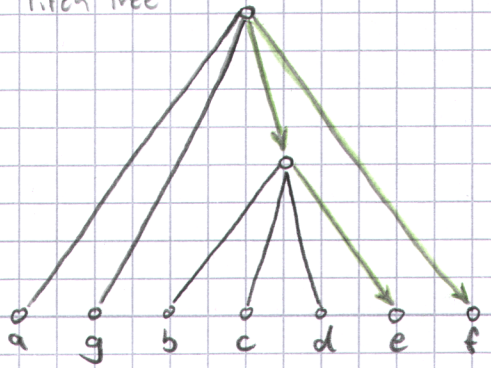


Reconstruction of event labeled gene trees

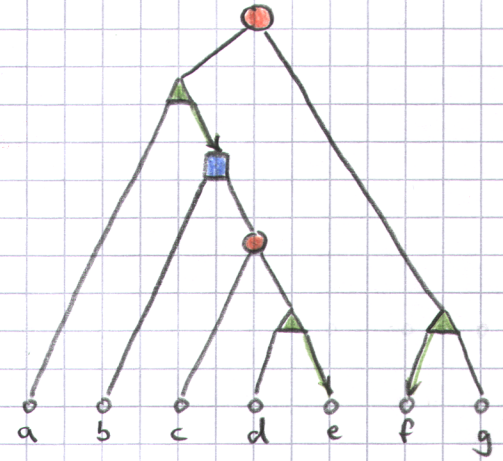


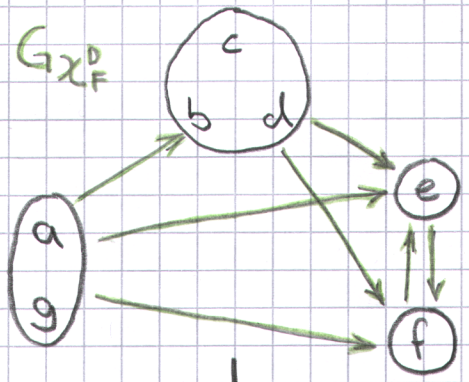


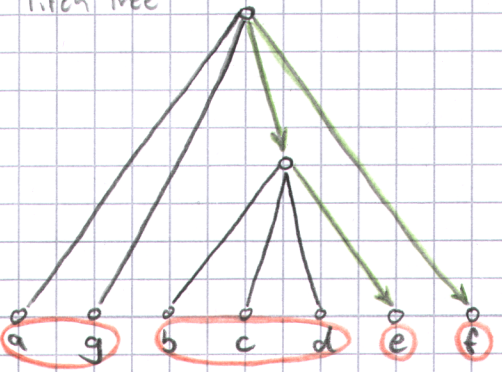
Fitch tree



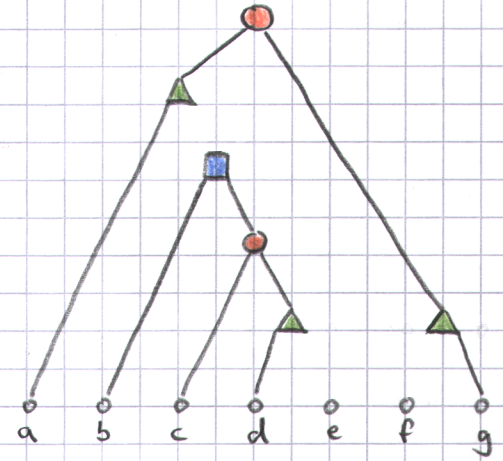
Reconstruction of event labeled gene trees

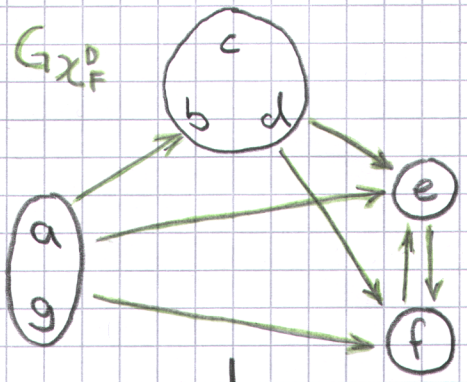


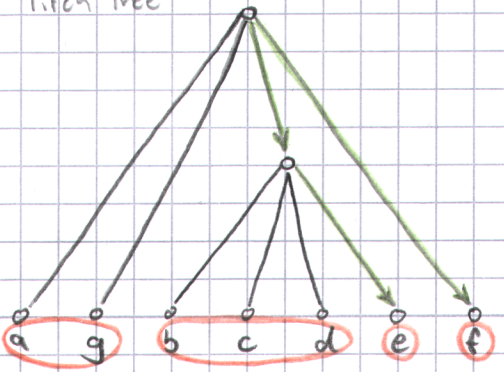




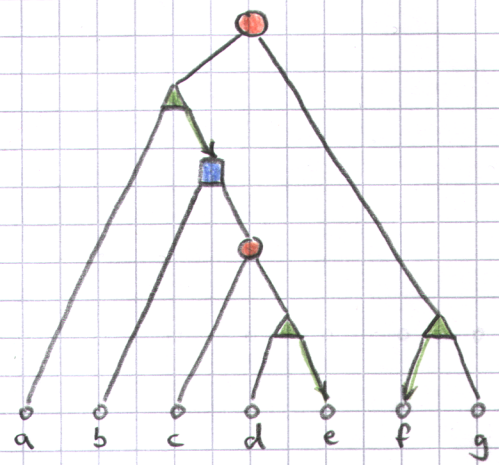
Reconstruction of event labeled gene trees



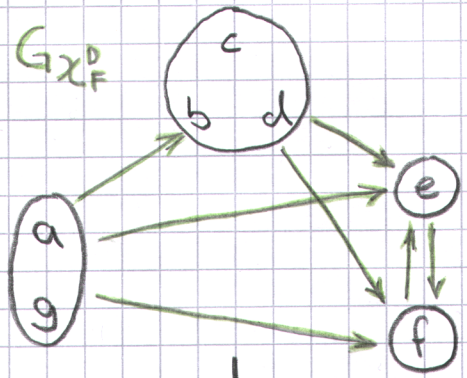




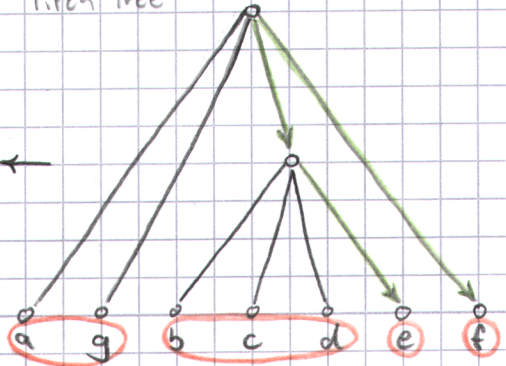
Reconstruction of event labeled gene trees

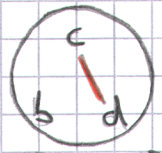


$G\chi_F^D$

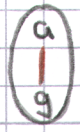


Fitch tree





$G_{O_F}[\{bucud\}]$

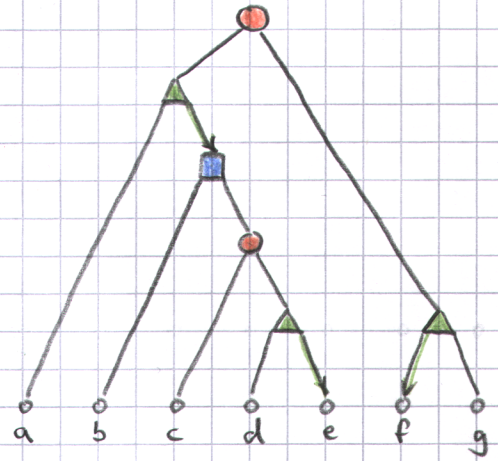


$G_{O_F}[\{aug\}]$

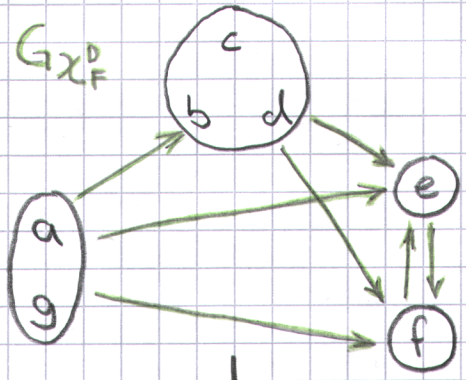
$G_{O_F}[\{e\}]$

$G_{O_F}[\{f\}]$

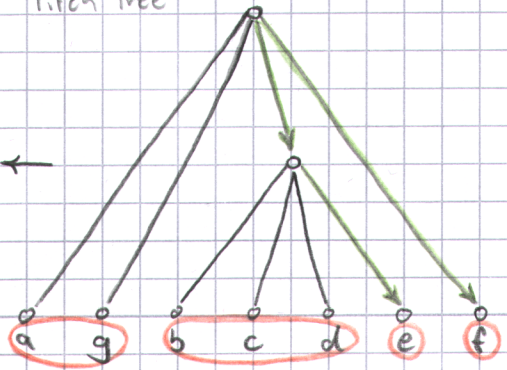
Reconstruction of event labeled gene trees

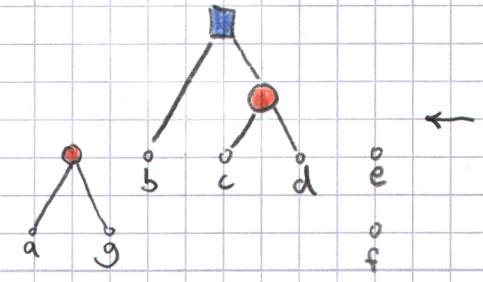


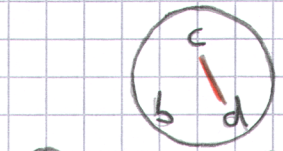
$G_{\chi_F^D}$



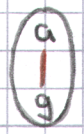
Fitch tree







$G_{\Theta_F}[\{bucud\}]$



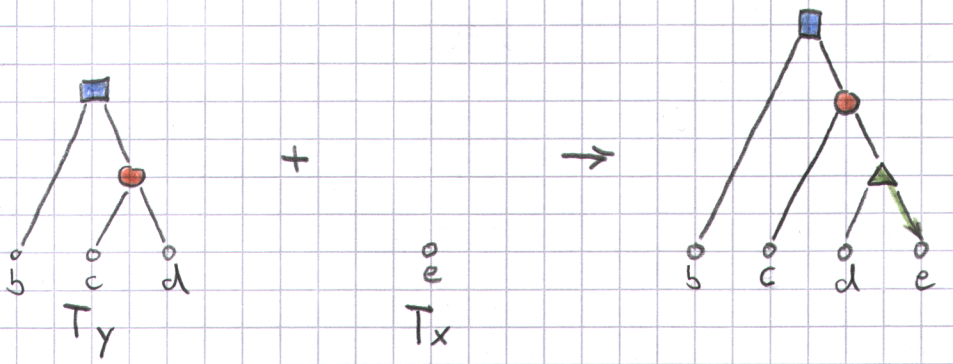
$G_{\Theta_F}[\{aug\}]$

$G_{\Theta_F}[\{e\}]$

$G_{\Theta_F}[\{f\}]$

Reconstruction of event labeled gene tree

How to append tree T_x to T_y



- find subtree $T_y(v)$ with $d(y,x) \ll d(y',x)$ for all $y \in L(T_y(v))$, $y' \in L(T_y) \setminus L(T_y(v))$ and $x \in L(T_x)$
- append T_x to T_y at edge (u,v) with HGT towards T_x

Summary

- improved orthology / xenology inference
- method for constructing event labeled gene trees from Fitch relations Θ_F, χ^D_F

Open Problems

- method relies on molecular clock-like data
- understand the connection between Θ_P, χ^D_F and species trees
- infer species trees
- construct time-consistent reconciliation map


Thank You...