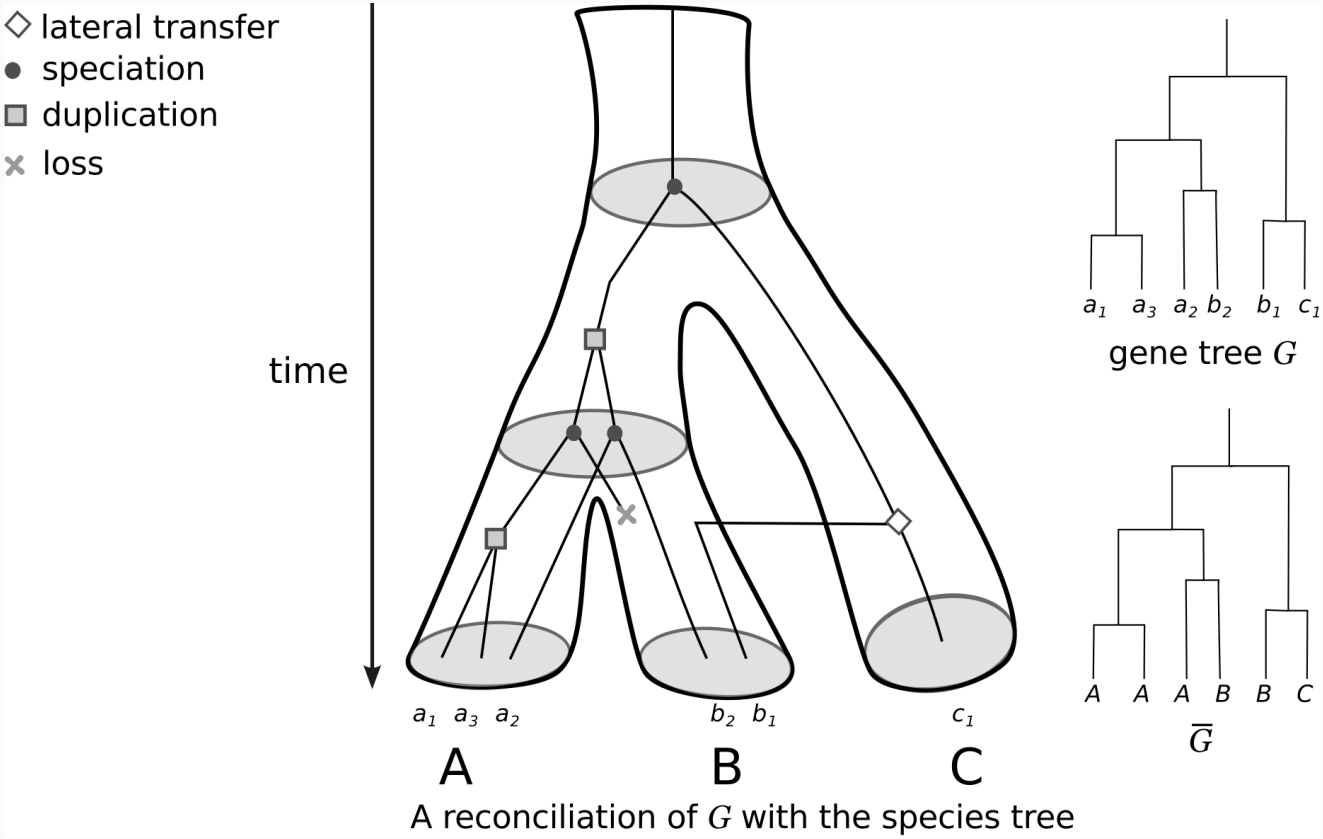
GATC: A Genetic Algorithm for gene Tree Construction under the Duplication Transfer Loss model of evolution

APBC 2018

Emmanuel Noutahi, Nadia El-Mabrouk
PhD candidate, UdeM, Canada

Université de Montréal

Gene family history



A reconciliation of G with the species tree

Gene family evolution

- Evolution by speciation, duplication, loss, HGT, etc
- Not necessarily observable from sequence data only.

“Including species tree information during gene tree reconstruction can tremendously improve accuracy”

Thomas [2010], Schreiber et al. [2013], Wu et al. [2013], Noutahi et al. [2016], etc

Integrative methods for gene tree reconstruction

Integrative methods usually include information from species tree through reconciliation between gene and species trees (explaning incongruence with gene gain and losses).

- Methods exploiting species tree topology information during the gene tree reconstruction process (GIGA, PhylDog, ALE, PrIME-dltrs, etc)
- *A posteriori* gene tree correction methods that explore alternative topology with better **reconciliation score** (ProfileNJ, Notung, Mowgli-NNI, TreeFix*, ecceTERA, etc)

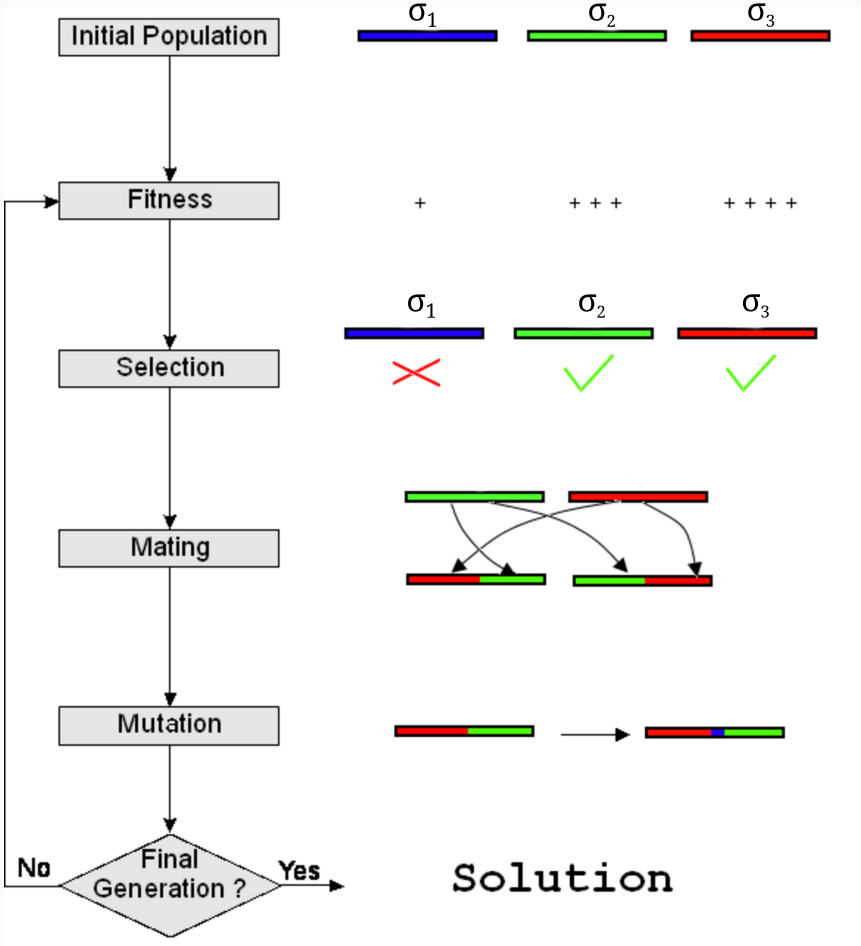
A new flexible approach ?

GATC: **G**enetic **A**lgorithm for gene **T**ree **C**onstruction/Correction Source code: `https://github.com/UdeM-LBIT/GATC`

Why GATC ?

- Most gene tree correction methods are incremental: sequence information can be lost during correction.
- Current integrative methods for construction use complex probabilistic models with many input requirements and have high computational cost (MCMC).
- ProfileNJ does not consider HGT events.
- Neighborhood exploration of only one tree increase the risk of being stuck in a local optimum.

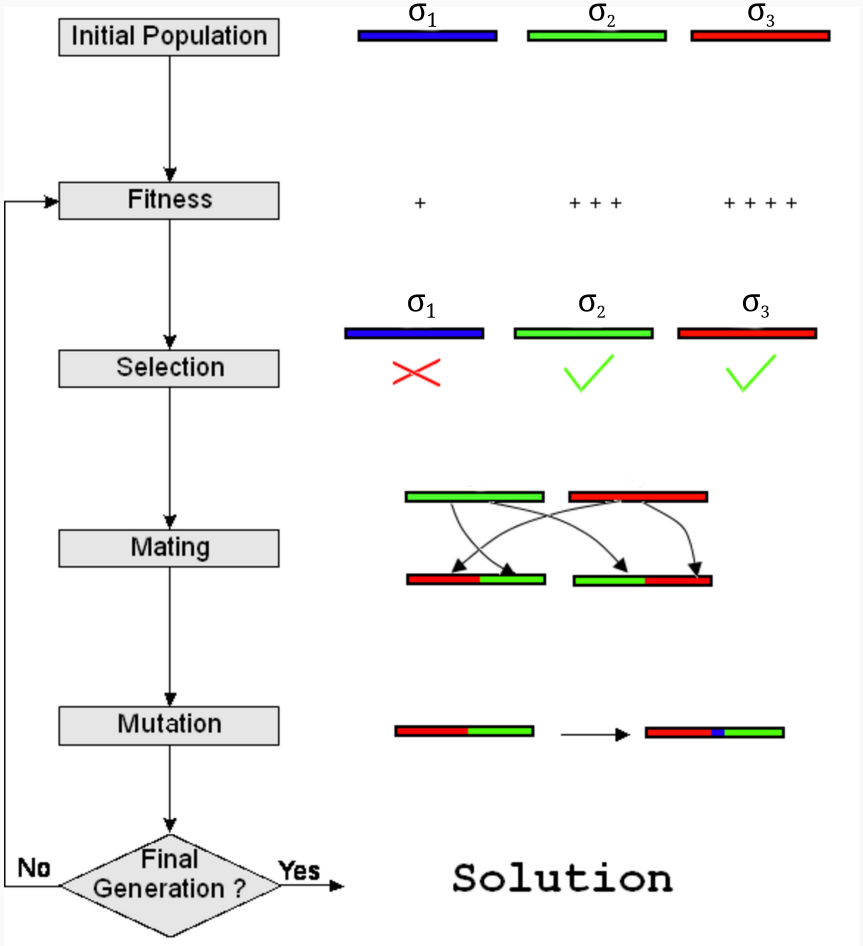
Genetic Algorithm



Generality

- Inspired by the process of "natural selection".
- Metaheuristic use for solving optimization problems
- some known application to tree reconstruction under ML (Matsuda [1996], Lewis [1998], Katoh et al. [2001], Zwickl [2006])

Genetic Algorithm



Notation

- Each *individual* has a *chromosome* σ_i which encode a specific solution to the problem.
- A *Population* of size n at *generation* k: $P_k = \{\sigma_i \mid 1 \leq i \leq n\}$
- *fitness function* evaluate the efficiency of each individual (*fitness score*) at solving the problem.

Encoding and initialization

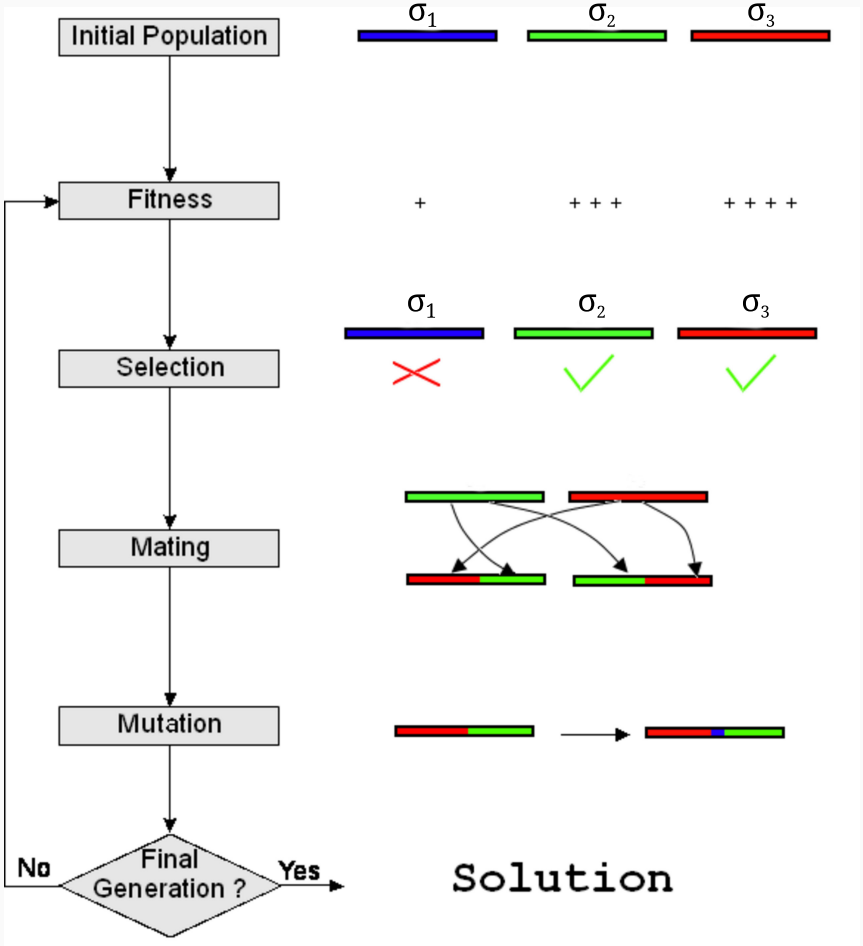
Solution encoding (σ_i)

- Each chromosome σ_i is encoded as $\left(G_i,\ \theta_i = \langle d, t, l, e, l, m \rangle\right)$
- Substitution model m is fixed for all generations from initialization.
- Event rates (d, t, l, e) can be fixed during evolution (depending on the reconciliation model).

Initial population (fixed size)

- Random trees
- bootstrap replicates
- **DL-only optimal solutions with PolytomySolver [Lafond et al., 2016] for ex.**
- Other tools output (gene tree correction mode).

Fitness computation: raw score vector $\vec{z}_i$



raw score $\vec{z}_i$ for each σ_i

- z_i^1: sequence likelihood score
- z_i^2: reconciliation score
- Optimization of z_i^1 only (classic ML tree construction)
  $\implies z_i^2 = 0, \forall\ i$

Computing sequence likelihood: z_i^1

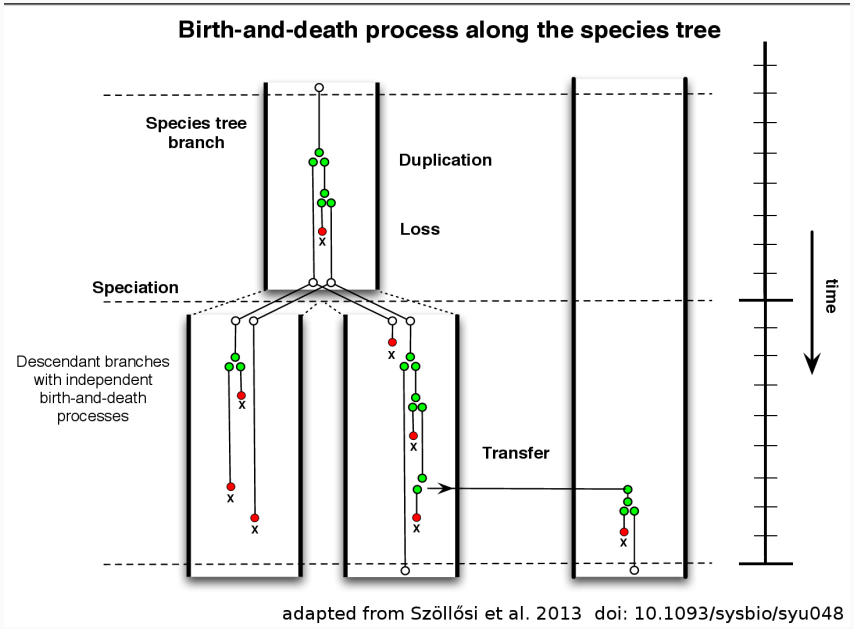
$z_i^1 = p(A|G_i, l_i, m)$: **sequence likelihood**

- Compute *likelihood* (and optimal model parameters) with Felsenstein algorithm [Felsenstein, 1981] (see RAxML [Stamatakis, 2014])
- Branch lengths l (re)estimated at this step.

Computing reconciliation score under DTL: z_i^2

Probabilistic models

- **DTLSR** [Tofigh, 2009], **ODT** [Szöllősi et al., 2012], **DLCoal** [Rasmussen and Kellis, 2012], etc
- Based on a Birth-Death and Gain model of gene family evolution.
- Need a discretization of the species tree into fixed time interval (dated species tree)
- $p(G|S, \lambda, \delta, \tau, e)$: integration on all possible reconciliation at each discretisation point (Slow !!).



Most Parsimonious Reconciliation

- Fixed event (duplication, transfer, loss) cost.
- Known polynomial-time algorithm (ex: **[Bansal et al., 2012]**), but NP-hard if time consistency is required [Tofigh et al., 2011].
- Works with undated trees.

Fitness f_i of an individual

Transformation into a single objective minimization problem

Let f_i denote the fitness of chromosome σ_i, $\vec{w}$ a weight vector and ϕ a scaling function.

$$f_i = \vec{w} \cdot \phi(\vec{z_i})$$

- $\approx$ joint likelihood $p(A|G, S)$ with probabilistic reconciliation
- ϕ : identity, sigmoid, etc
- $\vec{w}$: contribution of each component to overall fitness score.
- Does not make sense with MPR

Fitness under the MPR framework

Algorithm 1 Compute next generation population P_{k+1} from P_k

procedure COMPUTENEXTPOP(P_k)
  Compute P'_k, the offspring population of P_k
  $P^*_k \leftarrow P_k \cup P'_k$
  Evaluate z_i for all $\sigma_i \in P^*_k$
  Compute the dominance rank d_i for each $\sigma_i \in P^*_k$
  $w \leftarrow 1$
  while $\exists\ \sigma_i \in P^*_k \mid d_i = 0$ **do**
    $Wave_w \leftarrow \{\sigma_i \mid d_i = 0\}$
    Set a shared fitness for all $\sigma_i \in Wave_w$ as w
    $P^*_k \leftarrow P^*_k \setminus Wave_w$
    Compute the dominance rank d_i for each $\sigma_i \in P^*_k$
    $w \leftarrow w + 1$
  end while
  for $\sigma_i \in P^*_k$ **do**
    set the fitness of σ_i as $w + d_i$
  end for
  $P_{k+1} \leftarrow \bigcup_w Wave_w \cup P^*_k$
  return the first $|P_k|$ of P_{k+1} according to fitness
end procedure

Preliminary

- $\vec{z_i}$ is said to dominate $\vec{z_j}$, and noted $\vec{z_i} \prec \vec{z_j}$, iff $\vec{z_i} \neq \vec{z_j}$ and $z_i^1 < z_j^1$, $z_i^2 < z_j^2$
- Dominance rank $d_i = \sum_j a_{ij}$,

$$a_{ij} = \begin{cases} 1, & \text{if } \vec{z_j} \prec \vec{z_i} \\ 0, & \text{otherwise} \end{cases}$$

- Pareto Set (PS), set of non-dominated individuals : $PS = \{\sigma_i \mid \nexists\ \sigma_j, \vec{z_j} \prec \vec{z_i}\}$
- Main Hypothesis : PS_k correspond to the best solutions after k generations

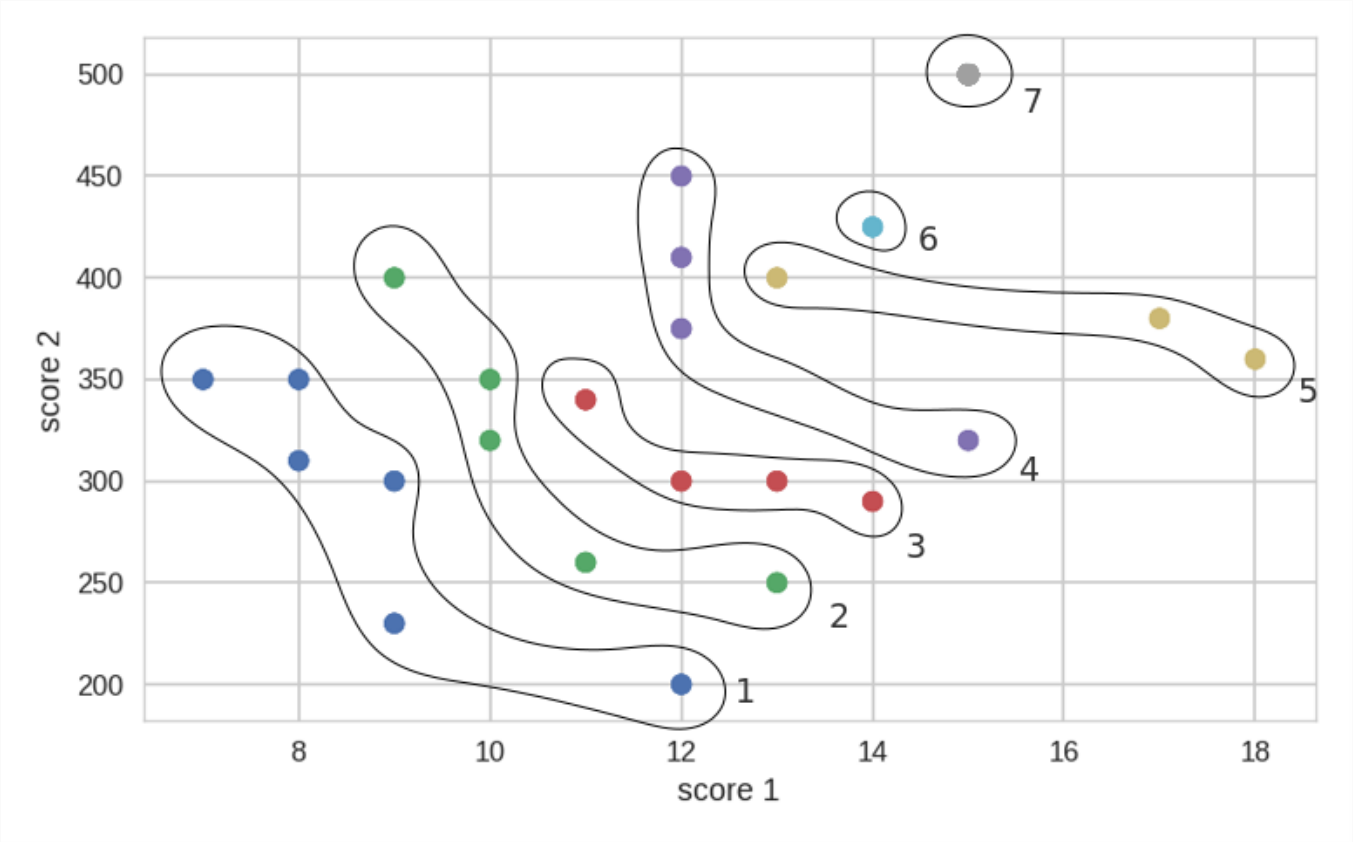
Fitness under the MPR framework

Algorithm 2 Compute next generation population P_{k+1} from P_k

procedure COMPUTENEXTPOP(P_k)
  Compute P'_k, the offspring population of P_k
  $P^*_k \leftarrow P_k \cup P'_k$
  Evaluate z_i for all $\sigma_i \in P^*_k$
  Compute the dominance rank d_i for each $\sigma_i \in P^*_k$
  $w \leftarrow 1$
  while $\exists\ \sigma_i \in P^*_k \mid d_i = 0$ **do**
    $Wave_w \leftarrow \{\sigma_i \mid d_i = 0\}$
    Set a shared fitness for all $\sigma_i \in Wave_w$ as w
    $P^*_k \leftarrow P^*_k \setminus Wave_w$
    Compute the dominance rank d_i for each $\sigma_i \in P^*_k$
    $w \leftarrow w + 1$
  end while
  for $\sigma_i \in P^*_k$ **do**
    set the fitness of σ_i as $w + d_i$
  end for
  $P_{k+1} \leftarrow \bigcup_w Wave_w \cup P^*_k$
  return the first $|P_k|$ of P_{k+1} according to fitness
end procedure

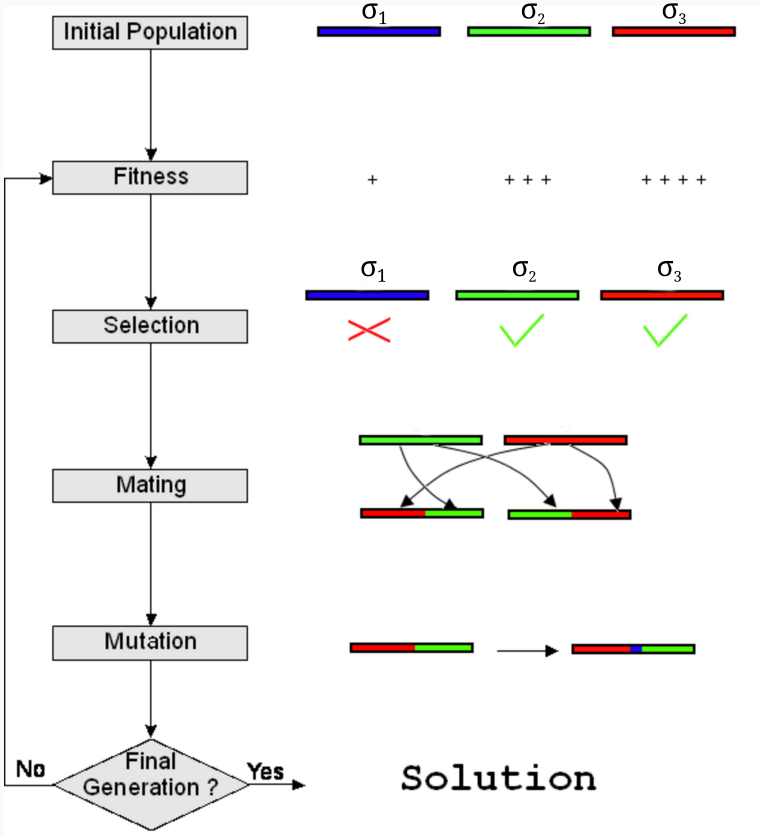
- Similar to the NSGA (Non-dominated Sorting Genetic Algorithm) for Multiple Objective Optimization Problems [Srinivas and Deb, 1994]
- Simultaneously consider **current population P_k and its offspring** $P_{k'}$ (after crossover + mutation)

Fitness computed by wave



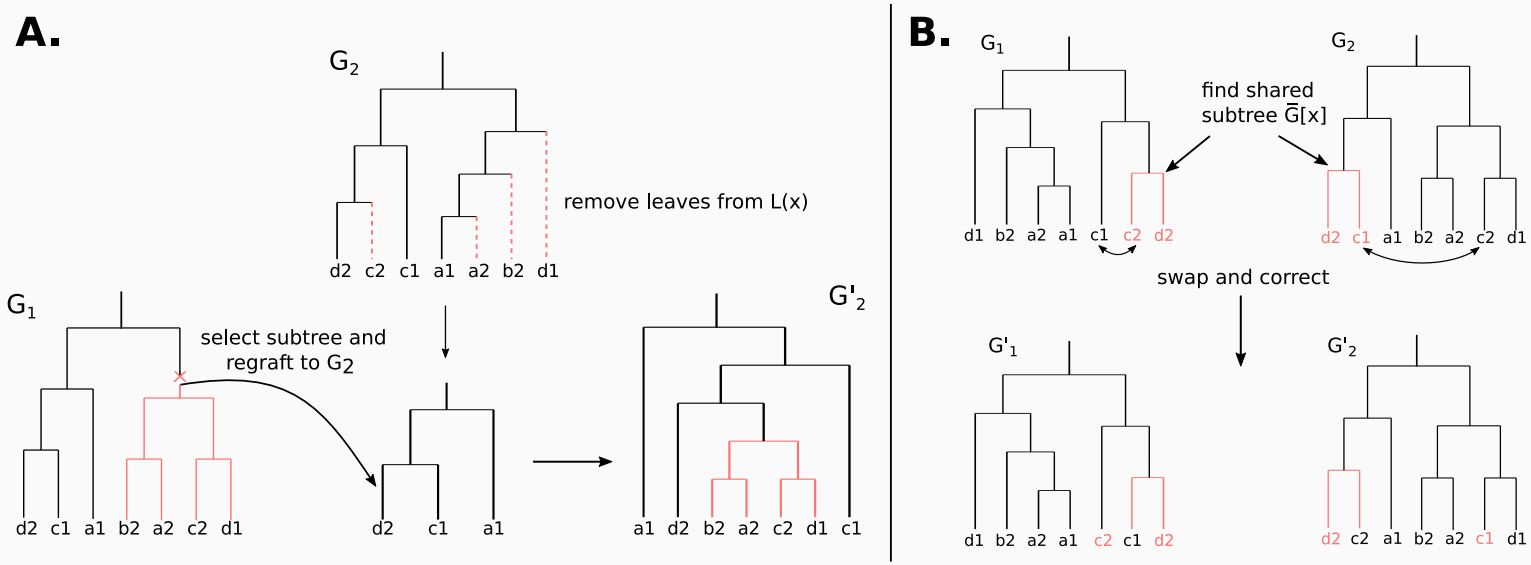
- Compute fitness in a *wave fashion* according to shared dominance rank.
- Best individuals have the lowest fitness value.

Selection process



- Random and uniform (ignore individual fitness)
- Roulette wheel : selection probability inversely proportional to fitness (recall that our best individuals have the lowest fitness value)
- Tournament selector : repeated roulette wheel on random subpopulation.

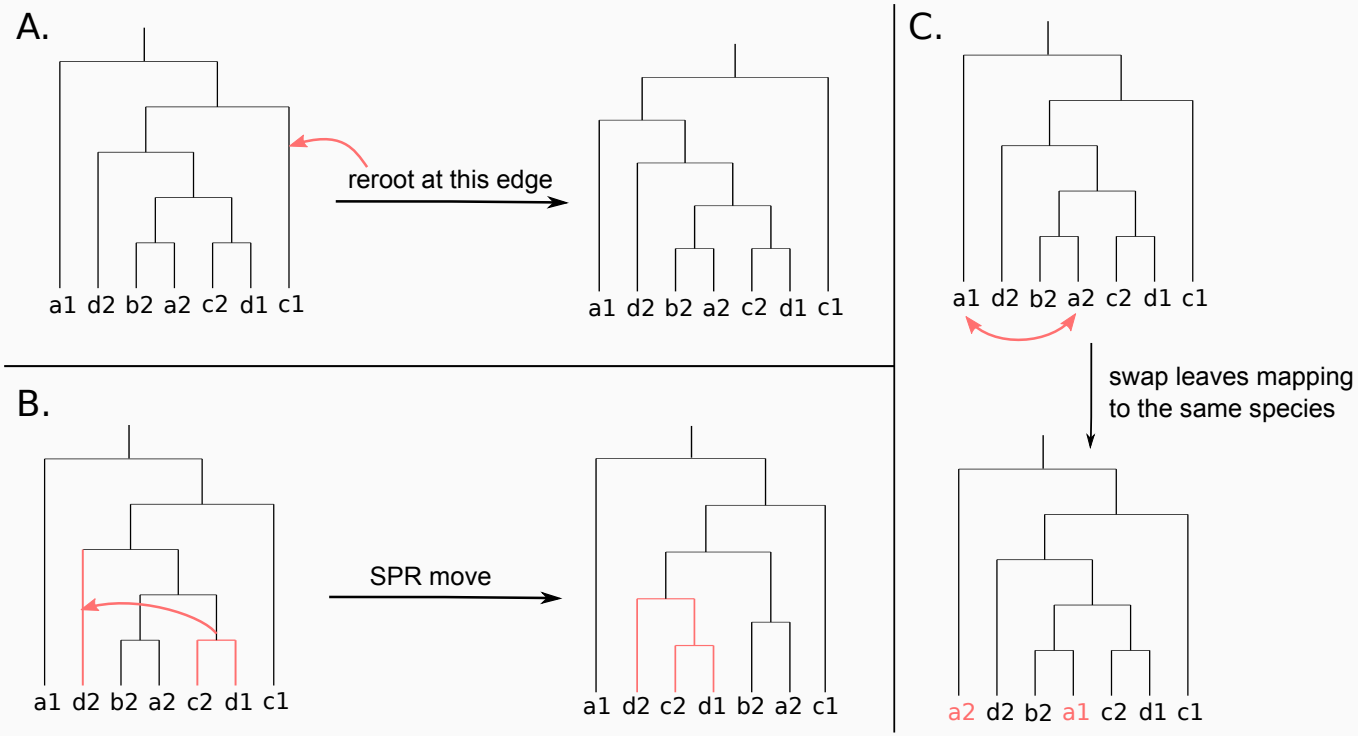
Mating (Crossover) between two chromosomes



A. Inter-chromosomal subtree swap. **B.** Swap preserving reconciliation score

- Only affect tree topology (G).
- Maintain population size (two children from selected pair of parent chromosomes).
- Input are trees with optimal reconciliation score $\implies$ leaf-labeling by gene problem
- Crossover rate P_{cross} control tree space exploration and influence the convergence of the GA.

Mutation



A. Re-rooting **B.** SPR move **C.** Leaf swap preserving reconciliation score

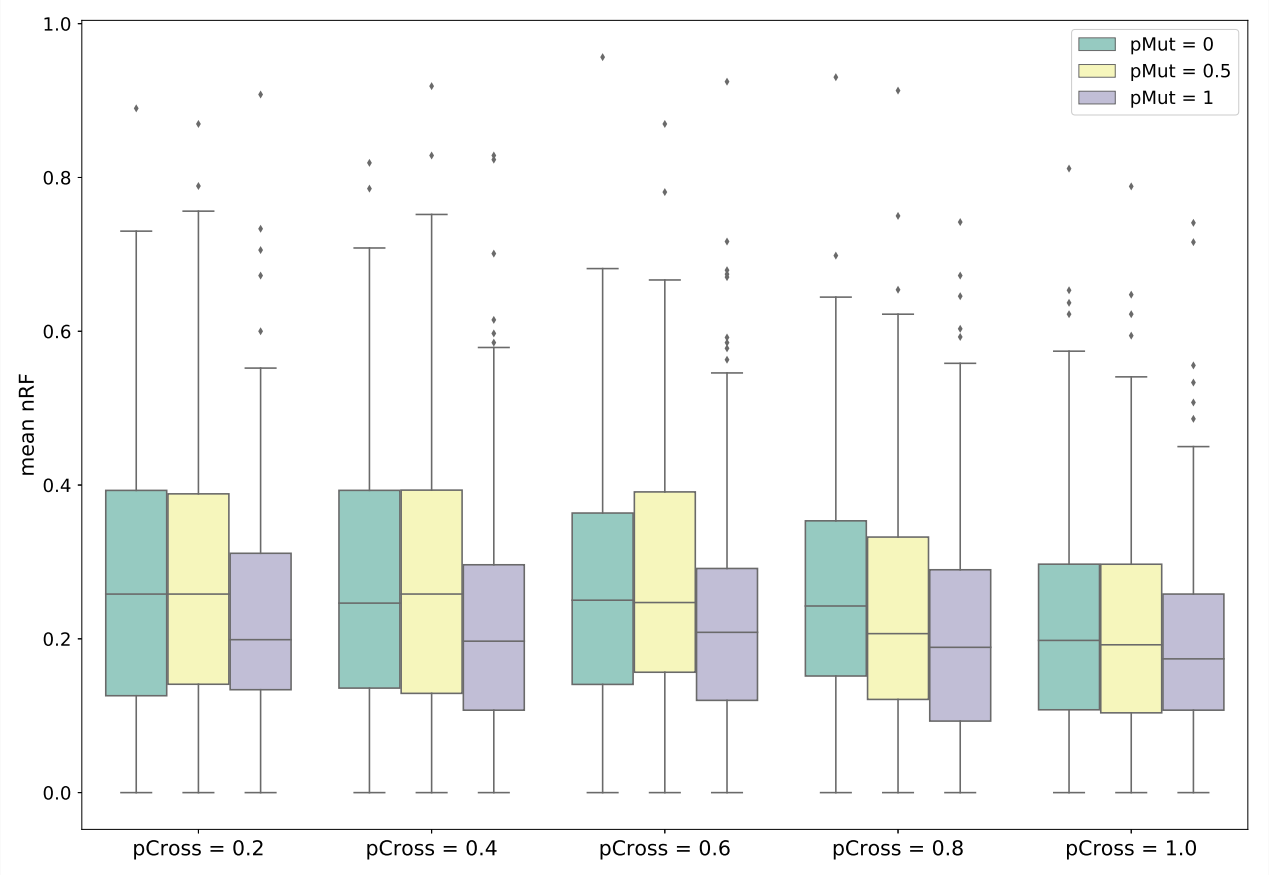
- Effect on tree topology, but other non-fixed parameters (d, t, l, e) can also be mutated (random sampling).
- Mutation add diversity and influence accuracy.

Stop criteria

Stopping criteria

- t_{max} : maximum evolution time
- n_{max} : maximum number of generation
- Convergence (unlikely)
  - $|PS| =$ popsize
  - popAU : statistically equivalent population according to AU test [Shimodaira and Hasegawa, 2001]
  - etc

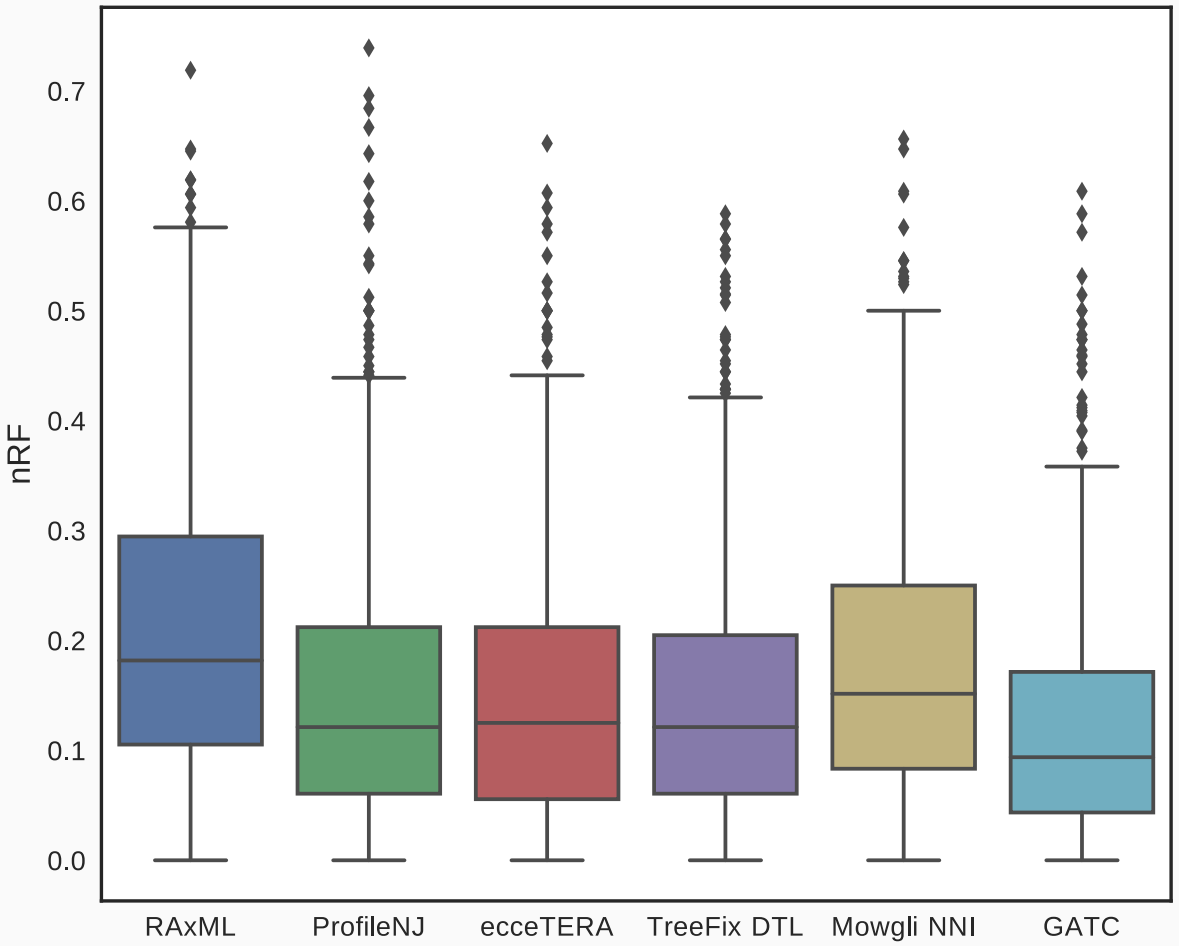
Influence of crossover and mutation rates on accuracy



Operator rates greatly influence accuracy

- Optimal rates will likely depends on dataset.
- Higher rates $\implies$ faster convergence, but risk of local optima. Balanced $(P_{cross}, P_{mut}) = (0.8, 0.5)$ as default.

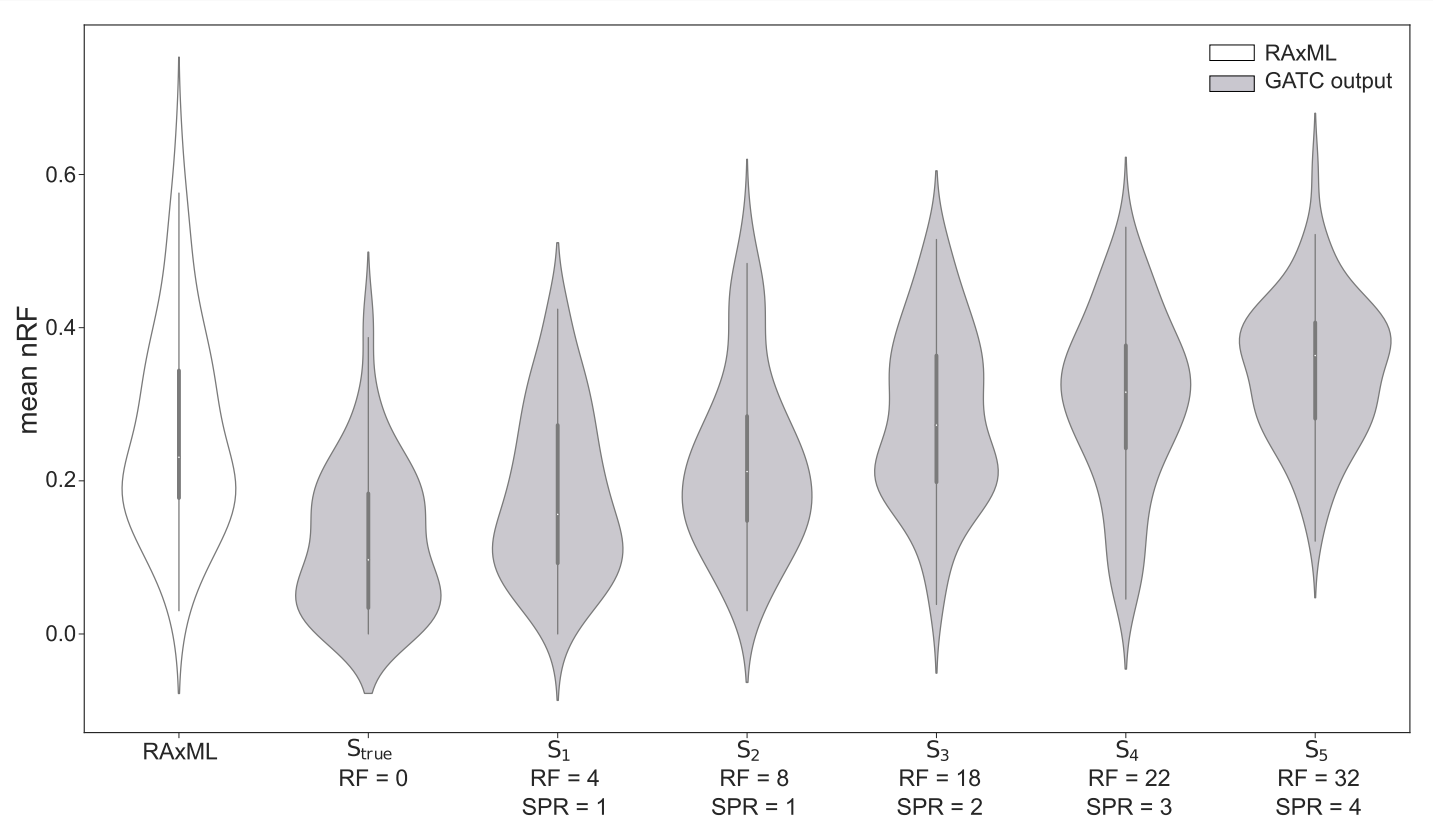
Results on simulated cyanobacteria dataset !



- Cyanobacteria dataset (1099 simulated alignments from ALE trees [Szöllősi et al., 2013]).
- Parameters: initialisation with 30 PolytomySolver trees, MPR with fixed event rates ($d = 2, t = 3, l = 1$) for reconciliation, LG + GAMMA, $P_{cross} = 0.8, P_{mut} = 0.5, t_{max} = 90$min, popAU stop criteria.

- Reconciliation-aware methods performed best.
- GATC outperformed other methods, but is much slower.

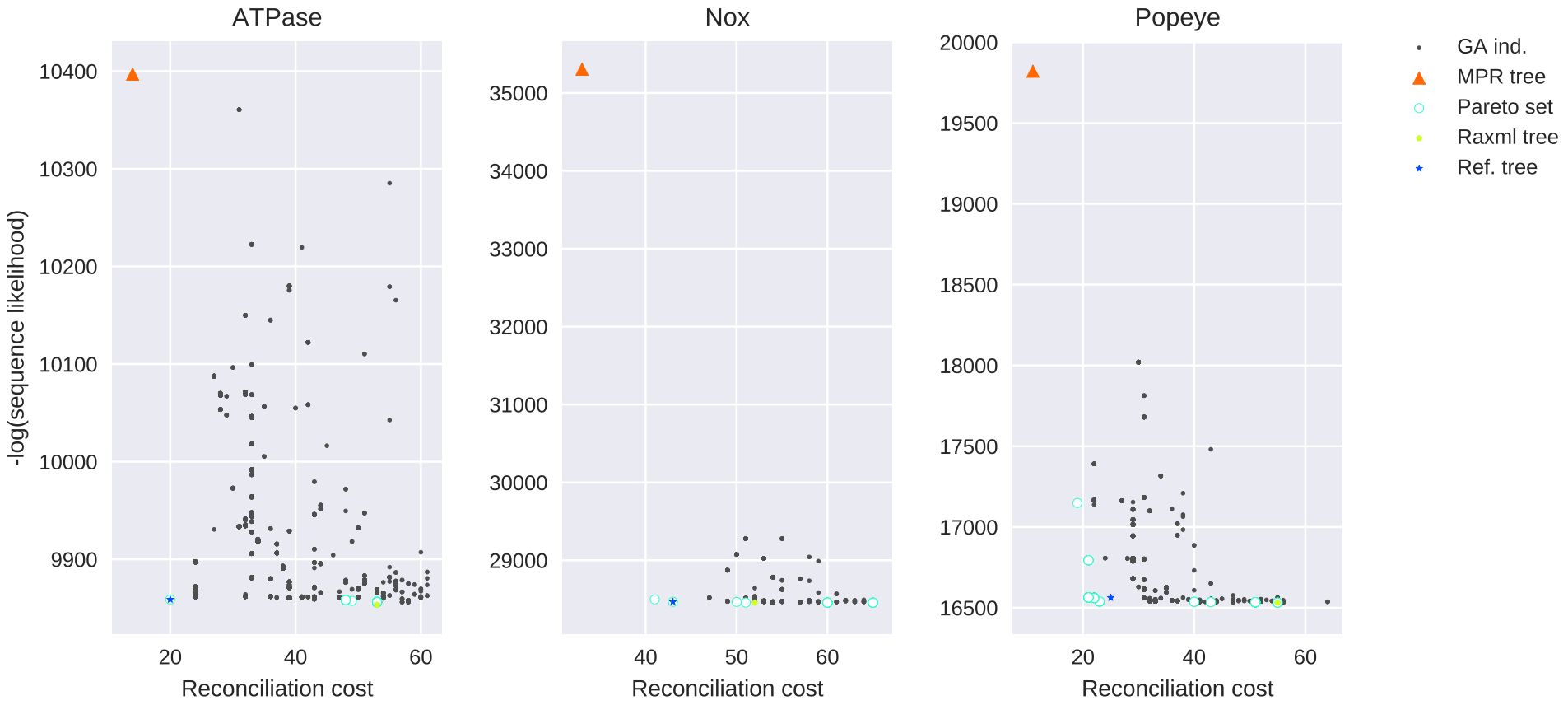
Limited effect of "alternative" species trees on accuracy



GATC seems robust to small topological errors in the species tree

- Performance measured by mean nRF score of last generation on 100 simulated trees: $(P_{cross}, P_{mut}) = (0.8, 0.5)$, default parameters, bootstrap replicates.
- Decreased accuracy with increasing errors, still performed better than RAxML for few errors.

GATC on 3 SwissTree reference trees



Reference tree mostly recovered in final Pareto Set

- "Gold Standard" trees for 3 eukaryotic protein families (manually obtained from the consensus of several methods [Boeckmann et al., 2011])
- GATC : DL only ($\tau = \infty$), $t_{max} = 3$h, $n_{max} = 300$

Popeye family case

Precision and Sensibility of inferred gene relationships

	normRF distance	Orthologs		Paralogs	
		Prec.	Rec.	Prec.	Rec.
Tree 1	0.260	0.763	0.942	0.971	0.871
Tree 2	0.260	0.765	0.941	0.971	0.873
Tree 3	0.087	0.902	0.983	0.992	0.894
Tree 4	0.109	0.829	0.866	0.940	0.922

- Similar topology
- GATC output have better likelihood and reconciliation score (fewer losses)
- GATC is a suitable for the construction of reference trees required for benchmarking gene tree construction softwares (time is not a limitation)

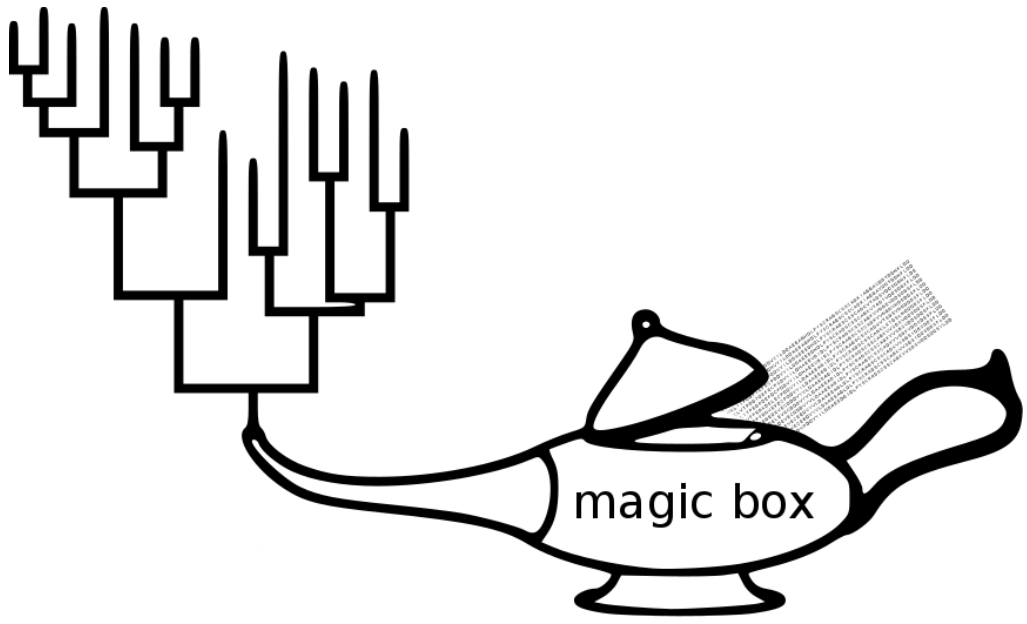
Going further

Current limitations	Workarounds
• No Incomplete Lineage Sorting	• Reconciliation framework can be extended to other events
• Unknown required evolution time for large trees	• **Parallel computing + Caching** , Efficient operators (ex: targeted SPR moves [Chaudhary et al., 2012])
• Multiple pareto optimal solutions	• **Sort by raw score**, **Amalgamation**, Filter with prior biological data.
• Optimal parameters (operator and DTL rates, initialisation algorithm, etc)	• Metapopulation with different settings and migration scheme.

Conclusion

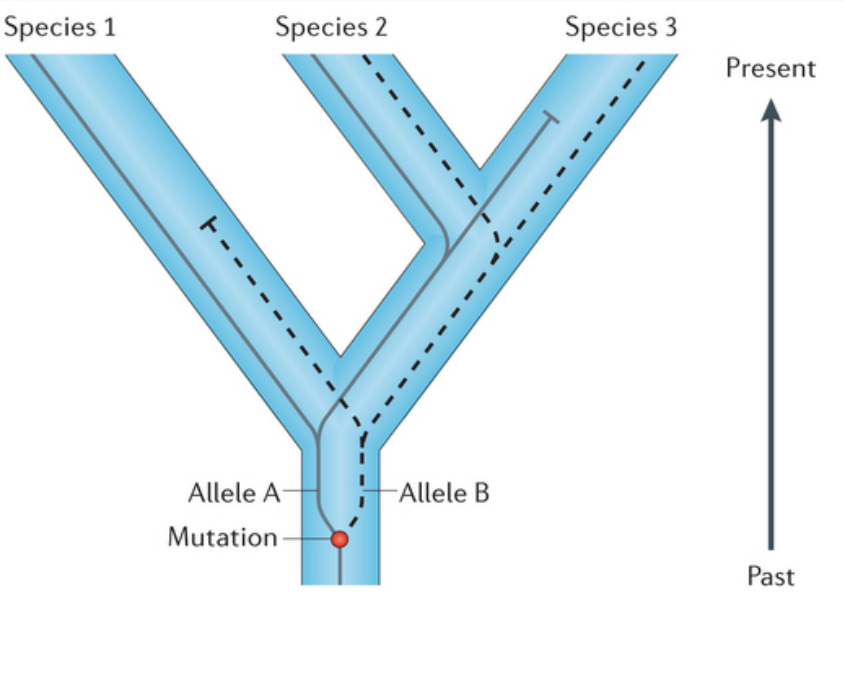
- Species tree aware methods often yield more accurate gene trees.
- Multiple Objective Optimization Algorithms are suitable for the *gene tree construction problem*.
- Pareto Set hypothesis seems to hold on both simulated and real data.
- Great alternative for the construction of reference trees.
- Room for improvement.

Questions ?



(GATC is freely available at `https://github.com/UdeM-LBIT/GATC` **)**

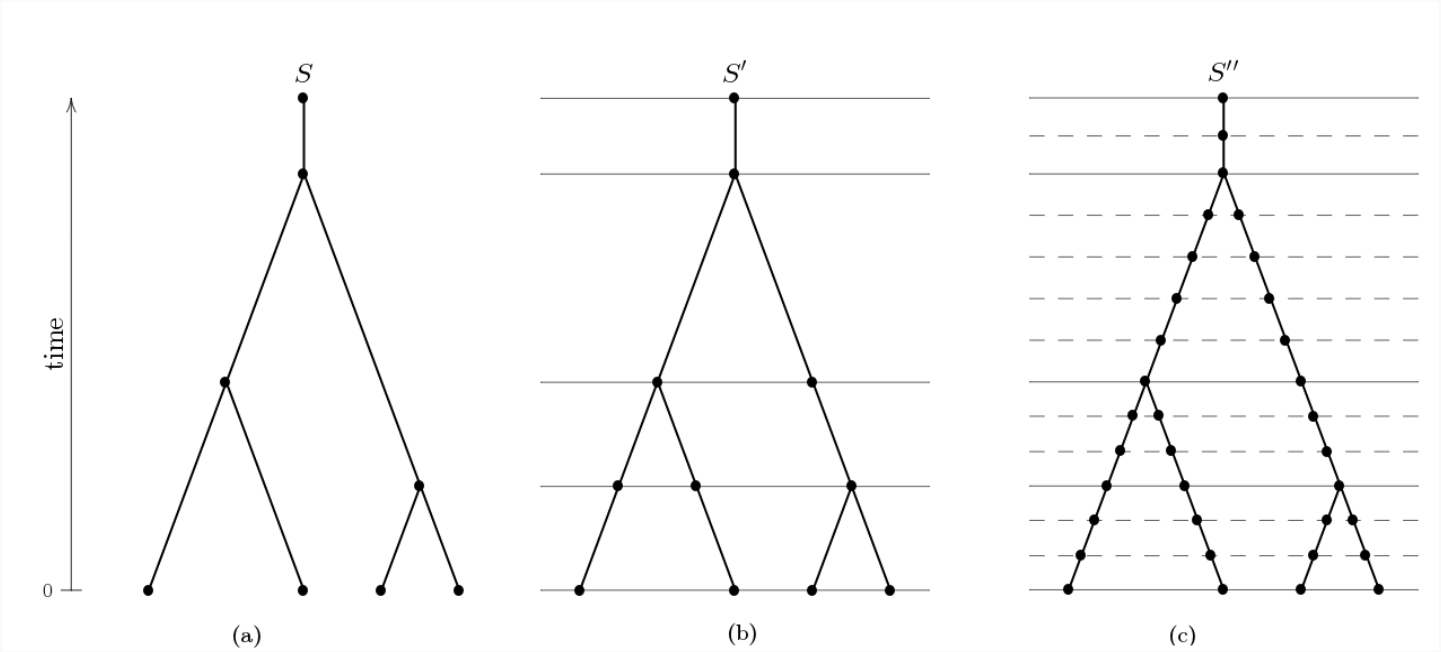
Incomplete Lineage Sorting (ILS)



ILS if allele B is lost in species 2.

Compute reconciliation likelihood (1/4)

Species tree discretization



- t_0 corresponds to extant species (leaves): we are going back in time.
- Nodes in S' correspond to speciation events.
- Discretization point in S'' correspond to possible event location.

Compute reconciliation likelihood (2/4)

Approximation de la vraisemblance

- $P(G, I|\theta) \approx \sum_{r \in \mathcal{R}} \sum_{d \in \mathcal{D}(r)} P(G, I, d|\theta)\Delta(d)$
- Sum over all possible reconciliations on all discretization nodes of the species tree.
- :ineage extinction probability

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_e(t) = \delta(Q_e(t))^2 + \tau\left(\sum_{f \in \mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|} Q_e(t)Q_f(t)\right) + \mu - \phi Q_e(t). \qquad (4)$$

For $e = \langle x, y\rangle \in E(S')$, the initial values for the system of equations above are given by

$$Q_e(t(y)) = \begin{cases} 0 & \text{if } y \text{ is a leaf,} \\ Q_f(t(y)) & \text{if } f \text{ is the single child of } e, \\ Q_f(t(y))Q_g(t(y)) & \text{if } f \text{ and } g \text{ are the two children of } e. \end{cases}$$

Tofigh et al. (2009)

Compute reconciliation likelihood (3/4)

Approximation de la vraisemblance

- Probability of single descendant from e to f

- f is contemporary to e

$$\frac{\mathrm{d}}{\mathrm{d}s}Q_{ef}(s,t) = 2\delta Q_e(s)Q_{ef}(s,t) - \phi Q_{ef}(s,t) + \tau \sum_{g\in\mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|}\Big(Q_{gf}(s,t)Q_e(s) + Q_{ef}(s,t)Q_g(s)\Big).$$

The initial values for the above equations are given by

$$Q_{ef}(t,t) = \begin{cases} 1 & \text{if } e = f, \\ 0 & \text{otherwise.} \end{cases}$$

- f is a descendant of e

$$Q_{ef}(s,t) = Q_{eg}(s,t(y))\Big(Q_{g'f}(t(y),t)Q_{g''}(t(y)) + Q_{g''f}(t(y),t)Q_{g'}(t(y))\Big) + \sum_{h\in\mathcal{C}_E(g)} Q_{eh}(s,t(y))Q_{hf}(t(y),t).$$

Tofigh et al. (2009)

Compute reconciliation likelihood (4/4)

Approximation de la vraisemblance

- Gene tree probability

$$p_{11}(e,x) = Q_{e'f'}(t(y),t(x)),$$

$$a(x,u) = s(e,v)s(f,w) + s(e,w)s(f,v),$$ x est un noeud de spéciation

$$a(x,u) = 2\delta s(e,v)s(e,w) + \tau \sum_{f \in \mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|}\Big(s(e,v)s(f,w) + s(e,w)s(f,v)\Big),$$ x entre deux noeuds de spéciations

$$s(e,u) = \begin{cases} p_{11}(e,\sigma(u))\rho\left(\frac{l(p(u),u)}{t(x)}\right) & \text{if } u \in L(G), \\ \sum_{z \in \mathcal{Q}(x)} p_{11}(e,z)\rho\left(\frac{l(p(u),u)}{t(x)-t(z)}\right) a(z,u) & \text{otherwise,} \end{cases}$$

Tofigh et al. (2009)

- Differential equations can be solved using Runge-Kutta method
- Can be optimized using matrix operations.

References

P.D. Thomas. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11:312, 2010.

F. Schreiber, M. Patricio, M. Muffato, M. Pignatelli, and A. Bateman. Treefam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Research*, 2013. doi: 10.1093/nar/gkt1055.

Y.C Wu, M.D. Rasmussen, M.S. Bansal, and M. Kellis. TreeFix: Statistically informed gene tree error correction using species trees. 62 (1):110- 120, 2013.

E. Noutahi, M. Semeria, M. Lafond, J. Seguin, L. Gueguen, N. El-Mabrouk, and E. Tannier. Efficient gene tree correction guided by genome evolution. *Plos.One*, 11(8), 2016.

References ii

H. Matsuda. Protein phylogenetic inference using maximum likelihood with a genetic algorithm. Pacific Symposium on Biocomputing, pages 512- 523, London, 1996. World Scientific.

P.O. Lewis. Genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.*, 15(3):277-283, 1998.

K. Katoh, K. Kuma, and T. Miyata. Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny. *J. Mol. Evol.*, 53:477- 484, 2001.

D.J. Zwickl. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, The University of Texas at Austin, 2006.

References iii

Manuel Lafond, Emmanuel Noutahi, and Nadia El-Mabrouk. Efficient non-binary gene tree resolution with weighted reconciliation cost. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 54. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313, 2014.

Ali Tofigh. *Using trees to capture reticulate evolution: lateral gene transfers and cancer progression*. PhD thesis, KTH, 2009.

References iv

Gergely J Szöllősi, Bastien Boussau, Sophie S Abby, Eric Tannier, and Vincent Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43):17513–17518, 2012.

Matthew D Rasmussen and Manolis Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, 22(4):755–765, 2012.

M.S. Bansal, J.A. Eric, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12):i283-i291, 2012. doi: 10.1093/bioinformatics/bts225.

Ali Tofigh, Michael Hallett, and Jens Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8 (2):517–535, 2011.

References v

Nidamarthi Srinivas and Kalyanmoy Deb. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3):221–248, 1994.

Hidetoshi Shimodaira and Masami Hasegawa. Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12): 1246–1247, 2001.

Gergely J Szöllősi, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. Efficient exploration of the space of reconciled gene trees. *Systematic biology*, 62(6):901–912, 2013.

Brigitte Boeckmann, Marc Robinson-Rechavi, Ioannis Xenarios, and Christophe Dessimoz. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Briefings in bioinformatics*, 12(5):423–435, 2011.

References vi

Ruchi Chaudhary, J Gordon Burleigh, and Oliver Eulenstein. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC bioinformatics*, 13(10):S11, 2012.