Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach

Cedric Chauve, Yann Ponty, João Paulo Pereira Zanetti

Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach

Cedric Chauve (SFU), Yann Ponty (CNRS/LIX/PIMS), João Paulo Pereira Zanetti (SFU/U. Campinas)

Background: DeCo

Reconstructing ancestral genome features is a classical comparative genomics problem, often addressed with dynamic programming (DP) algorithms. A DP algorithm, called **DeCo**, was recently introduced for computing **parsimonious** evolutionary scenarios for **gene adjacencies** in a **duplication-aware** framework, motivated by the reconstruction of **ancestral gene orders** [Bérard *et al., 2012*].

Results: DeClone

We describe **DeClone**, an extension of DeCo, using **Advanced Dynamic Programming** techniques, that allows the **sampling** of adjacency forests under the **Boltzmann distribution**, as well as the computation of **probabilities** of presence/absence of ancestral gene adjacencies, again under the Boltzmann distribution.

DeCo

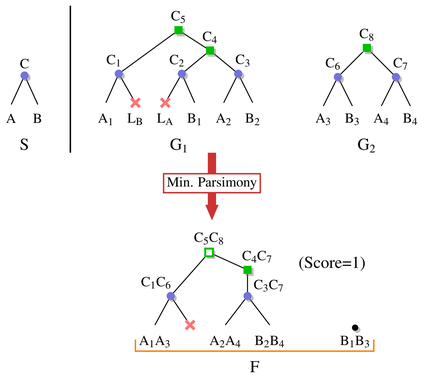
Input:

- Reconciled gene trees G_1 and G_2;
- Set of extant adjacencies.

Output:

- Max-parsimony adjacency forest

Parsimony: #adjacencies gains/breaks



Dynamic programming algorithm (excerpt):

$$c_1(v^1, v^2) = \min \begin{cases} \cdots \\ c_1(ca(v^1), cb(v^2)) + c_1(cb(v^1), ca(v^2)) \\ c_1(ca(v^1), cb(v^2)) + c_0(cb(v^1), ca(v^2)) + Break \\ c_0(ca(v^1), cb(v^2)) + c_1(cb(v^1), ca(v^2)) + Break \\ c_0(ca(v^1), cb(v^2)) + c_0(cb(v^1), ca(v^2)) + 2Break \\ c_1(ca(v^1), ca(v^2)) + c_1(cb(v^1), cb(v^2)) \\ c_1(ca(v^1), ca(v^2)) + c_0(cb(v^1), cb(v^2)) + Break \\ c_0(ca(v^1), ca(v^2)) + c_1(cb(v^1), cb(v^2)) + Break \\ c_0(ca(v^1), ca(v^2)) + c_0(cb(v^1), cb(v^2)) + 2Break \\ \cdots \end{cases}$$

Experiments

Data: $6,074$ DeCo instances, with genes taken from 36 mammalian genomes from the Ensembl database in 2012. **Syntenic inconsistencies** with parsimonious scenarios concern $5,817$, genes over $112,188$ ancestral and extant adjacencies.

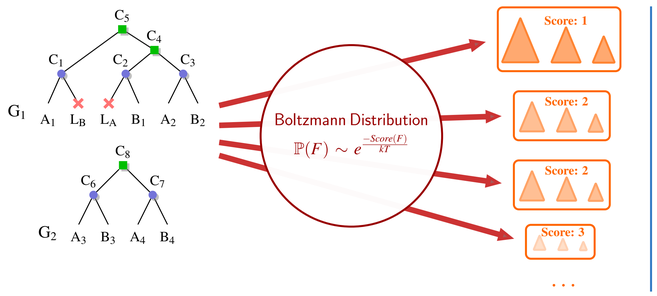
Methods: for each instance, we sampled $1,000$ adjacency forests under a Boltzmann distribution that favours highly co-optimal scenarios.

Results:

Adj. freq.	Anc. genes	Anc. adj.	Synt. Inc.
≥ 0.3	118,687	110,180	10,351
≥ 0.4	117,639	106,973	8,330
≥ 0.5	116,231	103,479	6,677
≥ 0.6	114,538	99,720	5,113
≥ 0.7	112,564	96,039	4,092
≥ 0.8	110,086	91,821	3,276
≥ 0.9	107,564	87,790	2,710
$= 1$	100,443	79,078	1,348

DeCo+Advanced Dynamic Programming=DeClone

The whole (sub)optimal space is modelled as a Boltzmann Ensemble

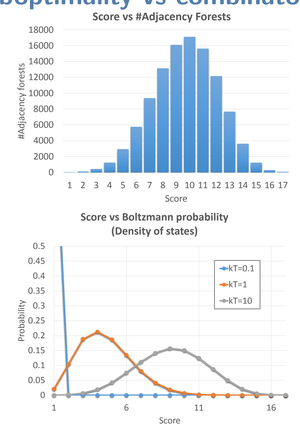


Analysis

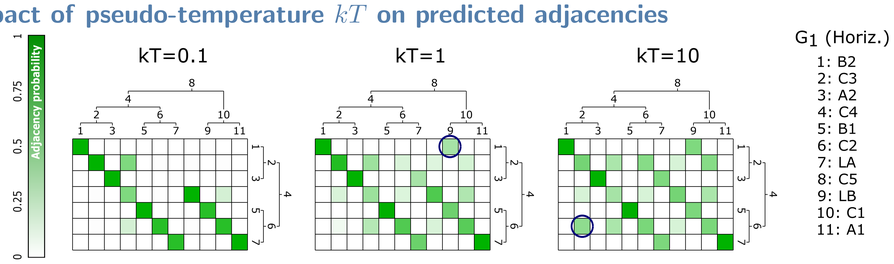
- Average properties
  - Ancestral adjacencies
  - Synthenic inconsistencies
  - Recurrent patterns ...
- Max. expected accuracy predictions
- Robustness assessment

...

Suboptimality vs combinatorics



Impact of pseudo-temperature kT on predicted adjacencies



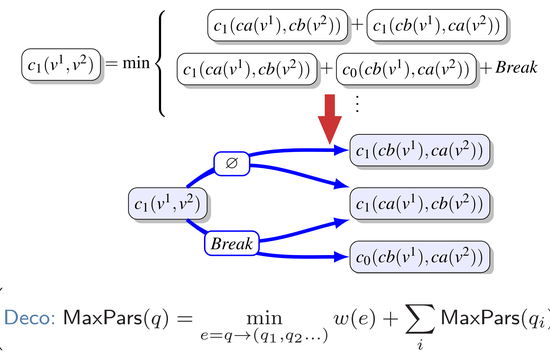
G_1 (Horiz.)	G_2 (Vert.)
1: B2	1: B4
2: C3	2: C7
3: A2	3: A4
4: C4	4: C8
5: B1	5: B3
6: C2	6: C6
7: LA	7: A3
8: C5	
9: LB	
10: C1	
11: A1	

Co-optimals (kT→0) — Temperature — **Uniform distribution** (kT→∞)

Algorithmic aspects

Algorithmic framework [Ponty/Saule, 2011]

Dyn. Prog. equations → Weighted Hypergraphs



(Deco: $\text{MaxPars}(q) = \min_{e=q\to(q_1,q_2\ldots)} w(e) + \sum_i \text{MaxPars}(q_i)$)

Claim: The DeCo DP scheme is **unambiguous** and **complete**

Algorithmic corollaries

- **Computing the partition function $\mathcal{Z}$ in $\mathcal{O}(|G_1||G_2|)$**
  Simple change of algebra $(\min, +, C) \to (\sum, \times, e^{-\frac{C}{kT}})$
- **Boltzmann sampling of adjacency forests**
  Each hyperedge chosen with probability proportional to its overall contribution to $\mathcal{Z}$.
- **Adjacency dot-plot:** $\mathcal{O}(|G_1||G_2|)$ inside/outside algorithm computes the probability of ancestral adjacencies.

Discussion

DeClone allows a controlled exploration of the space of all gene adjacency evolution scenarios: exhaustive enumeration, sampling, probabilities computation.
Experiments show that exploring the whole solution space under a Boltzmann distribution biased toward co-optimals allows to reduce significantly the number of syntenic inconsistencies observed when a single arbitrary optimal scenario is considered.

References

S. Bérard, *et al.*. 2012. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):382–388.
Y. Ponty and C. Saule. 2011. A Combinatorial Framework for Designing (Pseudoknotted) RNA Algorithms. *WABI 2011*, pp. 259–269.