# Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach

Cedric Chauve[1], Yann Ponty[1,2], and João Paulo Pereira Zanetti[1,3,4]

[1] Department of Mathematics, Simon Fraser University, Burnaby, Canada
[2] Pacific Institute for Mathematical Sciences, CNRS UMI3069, Vancouver, Canada
[3] Institute of Computing, UNICAMP, Campinas, Brazil
[4] São Paulo Research Foundation, FAPESP, São Paulo, Brazil

**Abstract.** We consider a recently introduced dynamic programming scheme to compute parsimonious evolutionary scenarios for gene adjacencies. We extend this scheme to sample evolutionary scenarios from the whole solution space under the Boltzmann distribution. We apply our algorithms to a dataset of mammalian gene trees and adjacencies, and observe a significant reduction of the number of syntenic inconsistencies observed in the resulting ancestral gene adjacencies.

## 1 Introduction

The reconstruction of the evolutionary history of genomic characters along a given species tree is a long-standing problem in computational biology. This problem has been well studied for several types of genomic characters, for which efficient algorithms exist to compute parsimonious evolutionary scenarios; classical examples include the case of genes and genomes sequences [9], gene content [5], and gene family evolution [1, 8]. Recently, Bérard *et al.* [3] extended the corpus of such results to syntenic characters, as they introduced the notion of adjacency tree, that models the evolution of gene adjacencies within a phylogeny, motivated by the reconstruction of the architecture of ancestral genomes and described an efficient dynamic programming (DP) algorithm, called `DeCo`, to compute parsimonious adjacency evolutionary histories.

From a methodological point of view, most existing algorithms to reconstruct evolutionary scenarios along a species tree in a parsimony framework rely on some dynamic-programming along this tree, whose introduction can be traced back to Sankoff in the 1970s for parsimony-based models (see [6] for a recent retrospective on this topic). Recently, several works have explored more general approaches for such parsimony problems that either explore a wider range of values for combinatorial parameters of parsimonious models [11, 10] or consider several alternate histories for a given instance, chosen for example from the set of all possible co-optimal scenarios (see [2, 14] for examples of this approach for the gene tree/species tree reconciliation problem), or from the whole solution space, thus including suboptimal solutions [7].

The present work follows the later approach and extends the `DeCo` DP scheme toward an exploration of the whole solution space of adjacency histories, under the Boltzmann probability distribution, that assigns a probability to each solution defined in terms of its parsimony score. This principle of exploring the solution space of a combinatorial optimization problem under the Boltzmann probability distribution, has been applied in several contexts and is sometimes known as the "Boltzmann ensemble approach" (see [13] for an illustration on RNA secondary structure prediction). While this Boltzmann ensemble approach has been used for a long time in RNA structure analysis, to the best of our knowledge it is not the case in comparative genomics, where exact probabilistic models have been favored as increasing computational capacities allow them to handle realistic datasets. However, such a probabilistic model does not exist so far for gene adjacencies, which motivates our work.

In the present paper, we modify the DP scheme of `DeCo` in order to sample gene adjacencies histories under the Boltzmann distribution. We apply our sampling algorithm to the dataset of over $6,000$ pairs of mammalian gene trees considered in the original `DeCo` paper [3]. This dataset was characterized by a significant number of ancestral genes involved in more than 2 adjacencies, which correspond to syntenic inconsistencies. We show that by sampling adjacencies histories under a Boltzmann distribution that favors co-optimal histories and conserving only frequent ancestral adjacencies, we can reduce significantly the number of syntenic inconsistencies.

## 2 Models

A *phylogeny* is a rooted tree which represents the evolutionary relationships of a set of elements represented by its nodes: internal nodes are ancestors, leaves are extant elements, and edges represent direct descents between parents and children. We consider here three kinds of phylogenies (illustrated in Figure 1): species trees, reconciled gene trees and adjacencies trees/forests. Trees we consider are always rooted. For a tree $T$ and a node $x$ of $T$, we denote by $T(x)$ the subtree rooted at $x$. If $x$ is an internal node, we assume it has either one child, denoted by $a(x)$, or two children, denoted by $a(x)$ and $b(x)$. A tree where all internal nodes have two children is called a *binary tree*.

*Species trees.* A species tree $S$ is a binary tree that describes the evolution of a set of related species, from a common ancestor (the root of the tree), through the mechanism of *speciation*. For our purpose, species are identified with *genomes*, and genes are arranged linearly or circularly along chromosomes.

*Reconciled gene trees.* A reconciled gene tree is also a binary tree that describes the evolution of a set of genes, called a *gene family*, through the evolutionary mechanisms of speciation, *gene duplication* and *gene loss*, within the given species tree $S$. Therefore, each node of a gene tree $G$ represents a gene loss, an extant gene or an ancestral gene. Ancestral genes are represented by the internal nodes of $G$, while gene losses and extant genes are represented by the leaves of $G$.

We denote by $s(g) \in S$ the species of a gene $g \in G$, and by $e(g)$ the evolutionary event that leads to the creation of the two children $a(g)$ and $b(g)$. If $g$ is an internal node of $G$, then $e(g)$ is a speciation (denoted by $\mathsf{Spec}$) if the species pair $\{s(a(g)), s(b(g))\}$ equals the species pair $\{a(s(g)), b(s(g))\}$, or a gene duplication ($\mathsf{GDup}$) if $s(a(g)) = s(b(g)) = s(g)$. Finally, if $g$ is a leaf, then $e(g)$ indicates either a gene loss ($\mathsf{GLoss}$) or an extant gene ($\mathsf{Extant}$), in which case $e(g)$ is not an evolutionary event.

*Adjacency trees and forests.* A *gene adjacency* is a pair of genes that appears consecutively along a chromosome. An adjacency tree represents the evolution of an ancestral adjacency through the evolutionary events of speciation, gene duplication, gene loss (these events, as described above, occur at the gene level and are modeled in the gene trees), and *adjacency duplication* ($\mathsf{ADup}$), *adjacency loss* ($\mathsf{ALoss}$) and *adjacency break* ($\mathsf{ABreak}$), that are adjacency-specific events.

- The duplication of an adjacency $\{g_1, g_2\}$ follows from the simultaneous duplication of both its genes $g_1$ and $g_2$ (with $s(g_1) = s(g_2)$ and $e(g_1) = e(g_2) = \mathsf{GDup}$), resulting in the creation of two distinct adjacencies each belonging to $\{a(g_1), b(g_1)\} \times \{a(g_2), b(g_2)\}$.
- An adjacency may disappear due to several events, such as the loss of exactly one (gene loss) or both (adjacency loss) of its genes, or a genome rearrangement that breaks the contiguity between the two genes (adjacency break).

Finally, to model the complement of an adjacency break, i.e. the creation of adjacencies through a genome rearrangement, *adjacency gain* ($\mathsf{AGain}$) events are also considered, and result in the creation of a new adjacency tree. It follows that the evolution of the adjacency between two genes can be described by a forest of adjacency trees, called an *adjacency forest*. In this forest, each node $v$ belongs to a species denoted by $s(v)$, and is associated an evolutionary event $e(v) \in \{\mathsf{Spec}, \mathsf{GDup}, \mathsf{ADup}\}$ if $g$ is an internal node, or $\{\mathsf{Extant}, \mathsf{GLoss}, \mathsf{ALoss}, \mathsf{ABreak}\}$ if $v$ is a leaf. Finally, adjacency gain events are associated to the roots of the trees of the adjacency forest.

*Parsimony scores and the Boltzmann distribution.* When considered in a parsimonious framework, the score of a reconciled gene tree $G$ can be either its duplication cost (number of gene duplications) or its mutation cost (number of gene duplications and losses). The score of an adjacency forest $F$ is the number of adjacency gains and breaks; other events are not considered as they are the by-products of evolutionary events already accounted for in the score of the reconciled gene trees $G_1$ and $G_2$. We denote by $s_a(F)$ the parsimony score of an adjacency forest $F$. Let $\mathcal{F}(G_1, G_2)$ be the set of all adjacency forests for $G_1$ and $G_2$, including both optimal and sub-optimal ones, where we assume that at least one extant adjacency is composed of extant genes from $G_1$ and $G_2$. The *partition function* associated to $G_1$ and $G_2$ is defined by

$$\mathcal{Z}(G_1, G_2) = \sum_{F \in \mathcal{F}(G_1, G_2)} e^{-\frac{s_a(F)}{kT}}$$

where $kT$ is an arbitrary constant. The partition function implicitly defines a *Boltzmann probability distribution* over $\mathcal{F}(G_1, G_2)$, where the probability of an

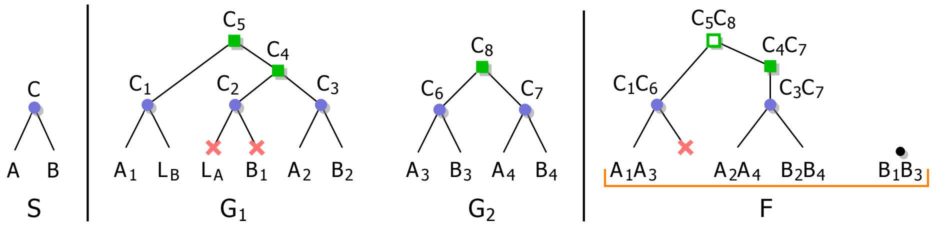


**Fig. 1.** (Left) Species tree $S$, with two extant species $A$ and $B$ and an ancestral species $C$. (Middle) Two reconciled gene trees $G_1$ and $G_2$, with four extant genes in genome $A$, four extant genes in genome $B$ and three ancestral genes in genome $C$. The set of extant adjacencies is $(A_1A_3, B_1B_3, A_2A_4, B_2B_4)$ (Right) Parsimonious adjacency forest $F$ composed of two adjacency trees. Blue dots are speciation nodes. Leaves are extant (species, genes, adjacencies), except when labeled with a red cross (gene loss). Green squares are (gene or adjacency) duplication nodes. Gene labels refer to the species of nodes. Every node of the adjacency tree is labeled by a couple of nodes from gene trees. Figure adapted from [3].

adjacency forest $F$ is defined by:

$$P(F) = \frac{e^{\frac{-s_a(F)}{kT}}}{\mathcal{Z}(G_1, G_2)}.$$

By exponentially favoring adjacency forests with lower parsimony scores, the Boltzmann distribution provides an alternative way to probe the search space, which is heavily influenced by the choice of $kT$. Indeed, decreasing $kT$ values will skew the Boltzmann distribution towards more parsimonious adjacency forests. Its limiting distributions are uniform over the whole search space ($kT \to +\infty$) or over the set of co-optimal forests ($kT \to 0$).

## 3 Algorithms

`DeCo`, the algorithm described in [3] to compute a parsimonious adjacency forest, is a DP scheme constrained by $S$, $G_1$ and $G_2$. We first present this algorithm, then describe how to turn it into a sampling algorithm.

*The* `DeCo` *DP scheme.* Let $G_1$ and $G_2$ be two reconciled gene trees and $g_1$ and $g_2$ be two nodes, respectively of $G_1$ and $G_2$, such that $s(g_1) = s(g_2)$. The `DeCo` algorithm computes, for every such pair of nodes $g_1$ and $g_2$, two quantities denoted by $c_1(g_1, g_2)$ and $c_0(g_1, g_2)$, that correspond respectively to the most parsimonious score of a parsimonious adjacency forest for the pairs of subtrees $G(g_1)$ and $G(g_2)$, under the hypothesis of a presence ($c_1$) or absence ($c_0$) of an ancestral adjacency between $g_1$ and $g_2$. As usual in DP along a species tree, the score of a parsimonious adjacency forest for $G_1$ and $G_2$ is given by $\min(c_1(r_1, r_2), c_0(r_1, r_2))$ where $r_1$ is the root of $G_1$ and $r_2$ the root of $G_2$.

So, $c_1(g_1, g_2)$ and $c_0(g_1, g_2)$ can be computed as the minimum of a sum of the scores of adjacency gains or breaks and, more importantly, of terms of the form $c_1(x, y)$ and $c_0(x, y)$ with $(x, y) \in \{g_1, a(g_1), b(g_1)\} \times \{g_2, a(g_2), b(g_2)\} - (g_1, g_2)$, using the two combinatorial operator min and $+$.

*(Un)-ambiguity of the* DeCo *DP scheme.* As defined in [13], the ambiguity of a DP algorithm can be defined as follows: a DP explores a combinatorial solution space (here for DeCo, the space of all possible adjacency forests, including possible suboptimal solutions), that can be explicitly generated by replacing in the equations min by $\Cup$ (the set-theoretic union operator) and $+$ by the Cartesian product $\times$ between combinatorial sets. A DP algorithm is then unambiguous if the unions are disjoint, i.e. the sets provided as its arguments do not overlap.

*Claim.* The DeCo dynamic programming scheme is unambiguous.

*Proof (Sketch).* Computing $c_1(g_1, g_2)$ and $c_0(g_1, g_2)$ branches on disjoint subcases that each involve a different set of terms $c_1(x, y)$ and $c_0(x, y)$. The only case that deserves a closer attention is the case where $e(g_1) = e(g_2) = \mathsf{GDup}$, as a simultaneous duplication can be obtained by two successive duplications. But in this case, the number of AGain events is different (we refer the reader to the original DeCo equations [3]), which ensures the obtained solutions are different.

*Stochastic backtrack algorithm through algebraic substitutions.* As mentioned in [13], any unambiguous dynamic programming scheme, and in particular that of DeCo, can be modified to not only exhaustively generate the set of all adjacency forests, but also to compute the corresponding partition function by replacing the arithmetic operators $(\min, +) \to (\sum, \times)$, and exponentiating atomic costs $C \to e^{-C/kT}$.

Also, and more importantly for our purpose, it allows to sample adjacency forests under the Boltzmann distribution, by changing the deterministic backtrack used for maximum parsimony into a stochastic operation. Indeed, assume that the partition function version of the DeCo equation computes $c_1(g_1, g_2)$ (resp. $c_0(g_1, g_2)$) as $\sum_{i \in [1,k_1]} t_i$, where the $t_i$ denote the contribution to the partition function of one of the local alternatives within the DP scheme. The latter are typically computed recursively as combinations of atomic adjacency gain/break costs, and *recursive* terms of the form $c_1(x, y)$ and $c_0(x, y)$ with $(x, y) \in \{g_1, a(g_1), b(g_1)\} \times \{g_2, a(g_2), b(g_2)\} - \{(g_1, g_2)\}$.

Then a (possibly non-parsimonious) solution can be built recursively for $c_1(g_1, g_2)$ (resp. $c_0(g_1, g_2)$), by branching on some $t_i$ with probability $t_i/c_1(g_1, g_2)$ (resp. $t_i/c_0(g_1, g_2)$), and proceed recursively on each occurrence of a *recursive* term within the alternative $t_i$. The correctness of the algorithm, i.e. the fact that the random processes induces a Boltzmann distribution on adjacency forests, follows readily from general considerations on unambiguous DP schemes [13].

The stochastic nature of the backtrack does not affect its worst-case complexity. Thus our Boltzmann sampling algorithm, for an instance composed of two gene trees $G_1$ and $G_2$ of respective sizes (number of leaves) $n_1$ and $n_2$, has time complexity of $\mathcal{O}(n_1 \times n_2)$ for each backtrack.

| Adjacency freq. | Ancestral genes | Ancestral adjacencies | Syntenic inconsistencies |
|---|---|---|---|
| $\geq 0.1$ | 120,269 / 122,367 | 116,204 / 136,480 | 14,851 / 31,221 |
| $\geq 0.2$ | 119,540 / 121,657 | 113,265 / 130,316 | 12,479 / 26,230 |
| $\geq 0.3$ | 118,687 / 120,631 | 110,180 / 121,307 | 10,351 / 19,369 |
| $\geq 0.4$ | 117,639 / 118,074 | 106,973 / 110,649 | 8,330 / 11,336 |
| $\geq 0.5$ | 116,231 / 112,812 | 103,479 / 99,672 | 6,677 / 5,522 |
| $\geq 0.6$ | 114,538 / 104,703 | 99,720 / 88,270 | 5,113 / 2,868 |
| $\geq 0.7$ | 112,564 / 92,449 | 96,039 / 74,482 | 4,092 / 1,256 |
| $\geq 0.8$ | 110,086 / 75,206 | 91,821 / 58,214 | 3,276 / 424 |
| $\geq 0.9$ | 107,564 / 45,702 | 87,790 / 33,648 | 2,710 / 33 |
| $= 1$ | 100,443 / 16 | 79,078 / 8 | 1,348 / 0 |

**Table 1.** Characteristics of ancestral genes and adjacencies from observed ancestral adjacencies filtered by frequency of observation (leftmost column), with values $kT = 0.1,\ 0.5$ (left / right).

## 4 Experiments

*Data.* We re-analyzed a dataset described in [3] composed of $5,039$ reconciled gene trees and $50,389$ extant gene adjacencies, forming $6,074$ DeCo instances, with genes taken from 36 mammalian genomes from the Ensembl database in 2012. In [3], these data were analyzed using the `DeCo` algorithm that computed a single parsimonious adjacency forest per instance. All together, these adjacency forests defined $112,188$ (resp. $96,482$) ancestral and extant genes (reps. adjacencies) and, more important, lead to $5,817$ ancestral genes participating to three or more ancestral adjacencies, which represent a significant level of syntenic inconsistency (close to 5%), as a gene can only be adjacent to at most two neighboring genes along a chromosome.

*Boltzmann sampling.* For each instance, we sampled $1,000$ adjacency forests under the Boltzmann distribution, for three values of $kT$, $0.001,\ 0.1,\ 0.5$. The results were very similar for values 0.1 and 0.001, so we will not discuss the later. The main difference between the results obtained with $kT = 0.1$ and $kT = 0.5$ is that in the later case, non-optimal adjacency forests have a higher probability in the Boltzmann distribution, and thus are more likely to be sampled, while $kT = 0.1$ skews the distribution toward optimal adjacency forests.

In order to assess the impact of considering alternate, possibly non-optimal, adjacency forests, we computed, for each ancestral adjacency observed in the sampled adjacency forests, its frequency, defined as the number of adjacency forests it is observed in divided by $1,000$, that is the number of sampled adjacency forests. Then, we considered 10 frequency thresholds, from 0.1 to 1 by steps of 0.1, and looked at the numbers of ancestral adjacencies, genes and syntenic inconsistencies from ancestral adjacencies whose frequency is at least the chosen threshold. Table 1 below summarizes our main observations.

*Discussion.* We can deduce two main points from the results summarized in Table 1. First, the difference observed between the results obtained in the two experiments clearly suggest that parsimony is an appropriate criterion for looking at gene adjacency evolution. Indeed, in the results obtained with $kT = 0.5$, that gives a higher probability to non-optimal adjacency forests, it appears that the number of conserved ancestral adjacencies drops sharply after frequency 0.6, showing that very few ancestral adjacencies appear with high frequency. Next, with $kT = 0.1$, we can observe that by taking a high frequency (starting at a frequency threshold of 0.6), we reduce significantly the number of syntenic inconsistency while maintaining a relatively similar number of ancestral genes than the experiments described in [3]; this observation is our main finding, and illustrates the interest of the ensemble approach compared to the classical dynamic approach that relies on a single arbitrary optimal solution.

## 5 Conclusion and perspectives

The main contribution of our work is an extension of the `DeCo` dynamic programming scheme to sample adjacency forests under the Boltzmann distribution. The application of our sampling algorithm on a mammalian genes dataset, together with a simple, threshold-based, approach to filter ancestral adjacencies, proved to be effective to reduce significantly the number of syntenic inconsistencies when sampling co-optimal adjacency forests, illustrating the interest of the ensemble approach. This preliminary work raises several questions and can be extended along several lines.

Sampling is obviously a non-exact way to estimate ancestral adjacency frequencies. Ideally, for each pair $(g_1, g_2)$ of ancestral genes from the same species in an instance, we would like to compute its exact frequency as an ancestral adjacency under the Boltzmann distribution, which corresponds to the Boltzmann probability of this feature. These probabilities can be computed exactly using a modification of the DP scheme, at no extra computational cost using the technique known as "inside/outside" algorithm. Similarly, the Boltzmann probabilities of the adjacency gains and breaks associated to ancestral adjacencies can be computed and then used to compute a *Maximum Expected Accuracy* adjacency forest, which is a parsimonious adjacency forest in a scoring model where each event is weighted by Boltzmann probability (see [4] for an example of this approach for RNA secondary structures).

Finally, we considered here an evolutionary model based on speciation, duplication and loss. A natural extension would be to include the event of lateral gene transfer in the model. Efficient reconciliation algorithms exist for several variants of this model [1, 8], together with an extension of `DeCo`, called `DeCo`LT [12]. `DeCo`LT is also based on dynamic programming, and it is likely that the techniques we developed in the present work also apply to this algorithm, and that the open question discussed above are of interest for this model.

## References


1. M. S. Bansal, E. J. Alm, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12):283–291, 2012.
2. M. S. Bansal, E. J. Alm, and M. Kellis. Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss. *Journal of Computational Biology*, 20(10):738–754, 2013.
3. S. Bérard, C. Gallien, B. Boussau, G. J. Szöllosi, V. Daubin, and E. Tannier. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18):382–388, 2012.
4. P. Clote, F. Lou, and W. A. Lorenz. Maximum expected accuracy structural neighbors of an RNA secondary structure. *BMC Bioinformatics*, 13(S-5):S6, 2012.
5. M. Csürös. Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In *RECOMB-CG*, volume 5267 of *Lecture Notes in Computer Science*, pages 72–86. Springer, 2008.
6. M. Csürös. How to infer ancestral genome features by parsimony: Dynamic programming over an evolutionary tree. In *Models and Algorithms for Genome Evolution*, pages 29–45. Springer, 2013.
7. J.-P. Doyon, S. Hamel, and C. Chauve. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(1):26–39, 2012.
8. J.-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. J. Szöllosi, V. Ranwez, and V. Berry. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *RECOMB-CG*, volume 6398 of *Lecture Notes in Computer Science*, pages 93–108. Springer, 2010.
9. D. A. Liberles, editor. *Ancestral Sequence Reconstruction*. Oxford University Press, 2007.
10. R. Libeskind-Hadas, Y.-C. Wu, M. S. Bansal, and M. Kellis. Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics*, 30(12):87–95, 2014.
11. L. Pachter and B. Sturmfels. Parametric inference for biological sequence analysis. *PNAS*, 101(46):16138–16143, 2004.
12. M. Patterson, G. J. Szöllosi, V. Daubin, and E. Tannier. Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(S-15):S4, 2013.
13. Y. Ponty and C. Saule. A combinatorial framework for designing (pseudoknotted) RNA algorithms. In *WABI*, volume 6833 of *Lecture Notes in Computer Science*, pages 250–269. Springer, 2011.
14. C. Scornavacca, W. Paprotny, V. Berry, and V. Ranwez. Representing a set of reconciliations in a compact way. *J. Bioinformatics and Computational Biology*, 11(2), 2013.