Error Detection and Correction of Gene Trees

Manuel Lafond * Krister M. Swenson † Nadia El-Mabrouk ‡

Abstract

Reconstructing the phylogeny of a gene family and reconciling the obtained gene tree with the species tree reveals the history of duplications, losses and other events that have shaped the gene family, with important implications towards the functional specificity of genes. However, evolutionary histories inferred by reconciliation are strongly dependent upon the accuracy of the trees, and few misplaced leaves will lead to a completely different history. Furthermore, sequence data alone often lack the information to confidently support a gene tree topology. We outline a number of criteria that can be used to detect erroneous gene trees. Analysing *Ensembl* gene trees of the fish genomes Stickleback, Medaka, Tetraodon, and Zebrafish reveals a significant number of erroneous gene trees. Finally, some potential directions for error correction of gene trees are explored.

1 Introduction

Duplication followed by modification is a major mechanism driving evolution. Consequently, genes cannot be seen as independent entities, but rather as entities related through duplication and speciation events. Grouping genes into families of *homologs* (*i.e.* copies originating from a single ancestral gene) and reconstructing the phylogeny of each gene family is requisite for a variety of annotation, evolutionary, and functional studies. By *reconciling* such a gene tree with a species tree, one can infer the history of duplications, losses and other events that have shaped the gene family. Such a history reveals the orthology (evolution of the ancestral copy by speciation) and paralogy (evolution by duplication) relationship between genes, with important implications towards the functional relationship between gene copies. However uncertainty on gene trees is a serious limitation to reconciliation, as well as to other applications. In particular, it has been reported that a few misplaced leaves can lead to a completely different history, possibly with significantly more duplications and losses [33]. Thus, a great deal of effort has been put into finding accurate gene trees.

Gene Tree Inference: Inferring phylogenies from sequence similarity is a field with a very long history that gave rise to a variety of distance, maximum parsimony, maximum likelihood or Bayesian methods, and a variety of software (PHYLIP [24, 25], NJ [49], PAUP [22], PhyML [31],

*Département d'informatique et de recherche opérationnelle (DIRO), Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, QC, H3C 3J7, Canada, lafonman@iro.umontreal.ca

†DIRO and McGill University, swensonk@iro.umontreal.ca

‡DIRO, mabrouk@iro.umontreal.ca

MrBayes [46], RAxML [52]). However, due to various limitations such as insufficient differentiation, alignment ambiguity, or differing rates of evolution among gene copies, sequences alone do not always support a single gene tree topology with high confidence.

Recently, several approaches have been developed to incorporate other genomic information in the construction of gene trees. For example, the SYNERGY algorithm [60] uses a "synteny similarity score" accounting for the position of genes in the chromosome. Different ways of integrating species tree information have also been considered. For example, the TreeBeST program from TreeFam [39, 47] (used for constructing the *Ensembl Compara* gene trees) uses a likelihood factor reflecting the number of duplications and losses inferred by reconciliation, the goal being to minimize inconsistency with the species tree. Another example is GIGA [57], a simple and fast algorithm using a UPGMA like distance-based approach to construct trees. In addition to the distance criterion, it relies on rules reflecting the species tree constraints (choose topologies in agreement with the species tree), as well as observations on lineage-specific evolution rates. This simple algorithm performs surprisingly well, leading to the conclusion that other constraints are strong enough to compensate for weak or misleading signals in gene sequences.

Other more sophisticated "species tree aware" methods have been developed, such as GSR [2, 1] and Spimap [45] adopting a Bayesian approach, or PhylDog [8] using a probabilistic model for simultaneously coestimating gene trees and the species tree. These models tend to be computationally intensive.

Gene Tree Correction: A complementary approach for producing "error-free" gene trees is to develop appropriate evaluation and correction tools, based on various genomic constraints, that can be applied subsequent to gene tree reconstruction. TreeFix [62] offers an additional framework to unify the sequence and genomic approaches, by suggesting a step following gene tree correction that performs statistical evaluation of a corrected tree, choosing it as a viable alternative only if it is statistically equivalent to the original one. The strategies that have been considered for gene tree correction are based on reconciliation, and can be grouped into three different classes:

I. Explore the space of gene trees obtained from the original one by performing some edit operations such as NNI [13, 28], SPR, or TBR [10] and select the tree having the minimum reconciliation cost. The "soft parsimony" algorithm [7] extends this approach for reconciliation with an uncertain species tree;

II. Collapse weakly supported internal branches [3], which leads to a non-binary gene tree, and then select the resolution minimizing the reconciliation cost [9, 38, 43];

III. Identify potentially misplaced leaves and remove them from the gene tree. In [12], vertices of a gene tree G labeled as Non-Apparent-Duplication (NAD) vertices, were flagged as potentially resulting from the misplacement of leaves in the gene tree. A duplication vertex x of G (according to the reconciliation with a given species tree) is a NAD if genes from the same species do not appear as a descendant of each of x's children. The reason for doubting NADs is that each one of these vertices reflects a phylogenetic incongruence with the species tree that is not due to the presence of duplicated genes in a single genome. Avoidance of NADs is one of the principles behind the GIGA algorithm [57]. We presented algorithmic results for removing, from a given gene tree, the minimum number of leaves or leaf-labels (species)

leading to a tree without a NAD vertex, under conditions of a known or an unknown species tree [16, 53]. All known formulations of this version of the problem are NP-hard [14, 15].

Error Detection: Known methods for correcting gene trees all rely on errors detected through reconciliation with the species tree. Similarly, in the field of gene tree reconstruction, most integrated methods rely on the species tree information, although other criteria have been suggested such as gene order [60] and variability of evolutionary rates [57]. In this paper, we follow up on this effort by exploring these two directions.

In Section 3, we show how gene order may be inconsistent with a gene tree, and state two error detection criteria based on gene order. To show the utility of these criteria, we consider the *Ensembl* [21] gene trees for four fish genomes (Stickleback, Medaka, Tetraodon, Zebrafish) with human and mouse as outgroups. We observe that more than 31% of all trees exhibit at least one gene order contradiction. In Section 4, we show how the presence of negative and positive selection may be misleading for gene tree reconstruction, and suggest methodology for detecting natural selection bias in a gene tree. Using the non-synonymous (dN) versus synonymous (dS) substitution ratio dN/dS as a criterion for detecting natural selection, a clear selective pressure is observed on Ensembl gene trees as compared to random trees. Finally, in Section 5 we give some avenues for developing a coherent tool for correcting gene trees, taking advantage of all available sequence and genomic information.

2 Genomes, Trees, and Gene Family Histories

We begin by introducing the necessary notations and background concepts. Although some of our experimental results could be explained without such formalities, we find it important to be precise. Indeed, many of the terms introduced in Section 2.5 have been used in multiple ways under diverse circumstances, sometimes leading to confusion. Many concepts are also presented in a general way, in the hopes of illuminating the potential for related work.

2.1 Genomes

Although our methods may be extended to arbitrary genomes, for simplicity of presentation we only consider single chromosomal genomes, represented as strings of, possibly signed, genes. Let $A = a_1 a_2 \cdots a_n$ be a string representing a genome. For any i, j such that $1 \leq i \leq j \leq n$, $A[i, j] = a_i a_{i+1} \cdots a_j$ is a substring of A. A string obtained from a substring of A by removing a subset of genes (possibly empty), is called a *subsequence* of A. For $1 \leq i_1 < i_2 < \cdots i_p \leq n$, we denote by $A[i_1, i_2, \cdots i_p]$ the subsequence $A[i_1]A[i_2] \cdots A[i_p]$ of A.

2.2 Trees

A *phylogeny* is a rooted binary tree, uniquely leaf-labeled by some set. A *species tree* S is a phylogeny over a set of species Σ, which represents the evolutionary relationships between these species. Similarly, we can consider the evolutionary relationships amongst a family of homologous genes Γ, that appear in the genomes of Σ. A *gene tree* G for Γ is a phylogeny accompanied by a function $s : \Gamma \to \Sigma$ indicating the species where each gene is found. We will make no

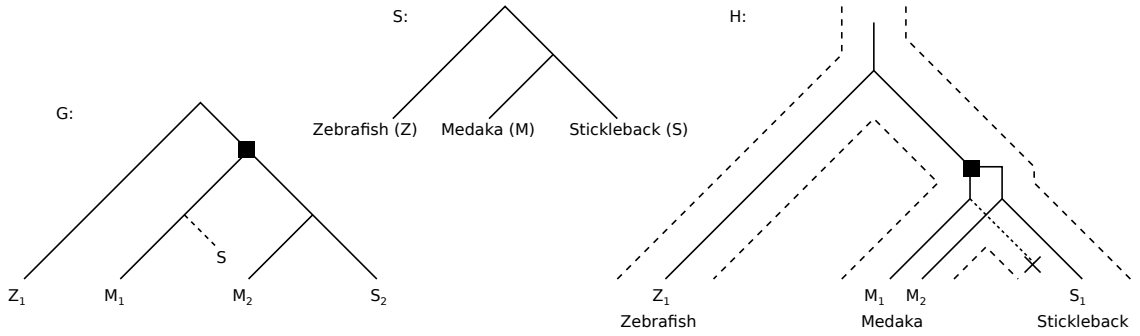


Figure 1: G is the gene tree for the gene family "OLA.11555" from Ensembl (ENSORLG00000013558), extended with a loss leaf according to a reconciliation with species tree S. H is the duplication-loss history corresponding to G. Reconciliation of G with the species tree S gives one duplication and one loss (as marked in G and H, duplication by a square, and losses by two dotted lines).

difference between a node and its associated gene. The tree G from Figure 1 is a gene tree for $\Gamma = \{Z_1, M_1, M_2, S_2\}$ on species set $\Sigma = \{Z, M, S\}$. In this case, $s(M_1) = s(M_2) = M$.

Given a tree T and a node x of T, we denote by T_x the subtree of T rooted at x (*i.e.* the tree comprises x and all its descendants), and by $\mathcal{L}(T_x)$ the set of leaves of T_x. The species set of x, denoted $\mathcal{S}(T_x)$, is the subset of Σ defined by the labels of the leaves of T_x (if T is a gene tree then $\mathcal{S}(T_x) = \{s(\ell) : \ell \in \mathcal{L}(T_x)\}$). If there is no ambiguity about the tree in question, we write $\mathcal{S}(T_x)$ as $\mathcal{S}(x)$. The *lowest common ancestor* (LCA) of leaves x and y in a tree T, written $lca_T(x, y)$, is the common ancestor of x and y that is farthest from the root. Finally, for any internal node x of a rooted binary tree T, we denote by x_ℓ and x_r the two children (left and right) of x in T.

2.3 Histories

As a set of modern species evolves from a single ancestral species, some of the gene content of those species is modified through duplication within the genome, and then loss. Traditionally, reconciliation between gene trees and species trees has been used to reconstruct such histories. The basis for such methodology has been a formal definition of what a reconciliation is, without a definition of the actual history that is the ultimate objective. Indeed, for a family of genes related through duplications and speciations, there exists some true history – the actual duplications and speciations that occurred in the past. So as not to put the cart in front of the horse, we now define what we mean by a *duplication/loss/speciation-history* (*dls*-history). We then define a *reconciliation* in terms of *dls*-histories. This perspective facilitates the reasoning used in Section 3; knowing that there is one true history of speciation/loss/duplication for a family of genes, we establish conditions that true gene trees must possess.

A *duplication* of size $k+1$ on genome A is an operation that copies a substrings $A[i, i+k]$ to a location j of A outside the interval $[i, i+k]$ (*i.e.* preceding i or following $i+k$). A *Loss* of size k is an operation that removes a substring of size k from A. Given a set of genes Γ from a set of genomes Σ, a *duplication/loss/speciation-history* H for Γ is a rooted tree "embedded" in the species tree S of Σ, which reflects the evolution of the set from a single ancestral copy through

duplication, loss and speciation events. In other words, each internal node x of H represents the evolution of the set $\mathcal{L}(H_x)$ from an ancestral gene copy x_A, and corresponds to either a speciation or gene duplication event. The leaves correspond to either the genes in question, or to losses, where each of the latter *loss leaves* map to a single node of S. If a loss leaf ℓ maps to a node x of S, we say that S_x is the *label* of ℓ.

Definition 1 (*dls*-history) *Let Γ be a set of genes from a set of genomes Σ, and let S be the true phylogeny for Σ. A* duplication/loss/speciation-history *H for Γ consistent with S (or simply a* dls*-history if unambiguous) is a rooted binary tree such that:*

- *each leaf is uniquely labeled by an element of Γ, or it is a loss leaf labeled by a subtree of S;*
- *each internal node is labeled as a duplication or speciation; and*
- *H is* consistent *with S: Consider the tree $\bar{H}$ obtained from H by replacing each loss leaf by the subtree that labels it, and by replacing all other leaves by the species to which the attached gene belongs. Then, for every internal node x of $\bar{H}$ such that $|\mathcal{S}(x)| \geq 2$, there exists a vertex u of S such that $\mathcal{S}(x) = \mathcal{S}(u)$ and: $\mathcal{S}(x_r) = \mathcal{S}(x_\ell)$ if x is a duplication, or $\mathcal{S}(x_r) = \mathcal{S}(u_r)$ and $\mathcal{S}(x_\ell) = \mathcal{S}(u_\ell)$ if x is a speciation node.*

The gene tree *in agreement* with H is the tree obtained from H by removing loss leaves and the resulting internal nodes having one child. Consider the trees from Figure 1. The solid lines of G denote the gene tree corresponding to the history H.

As true histories are unknown, gene trees are usually inferred from sequence data, and histories subsequently inferred from *reconciliation* with the species tree (see the next section). In this paper, we will distinguish between the *true gene tree*, which is the tree in agreement with the true *dls*-history of the gene family, and the *gene tree*, which is a tree obtained from the observed gene sequences (*e.g.* a multiple alignment of the sequences, the observed gene positions, or any other footprint of evolution observed in the extant species).

2.4 Reconciliation

Given an inferred gene tree G for a set Γ of genes from genomes Σ, and given a species tree S for Σ, the problem is to recover a *dls*-history for Γ consistent with S, such that G is in agreement with the history. Such a history is called a *reconciliation*. Informally, a *reconciliation* R of G and S is a *dls*-history of Γ obtained by inserting loss leaves in G. Let an *extension* of G be a tree obtained from G by a sequence of loss insertions, where a *loss insertion* denotes the insertion of a new loss leaf labeled by a subtree of S, by means of bisecting an existing edge of G with a new edge. A rigorous definition of reconciliation follows.

Definition 2 (reconciliation) *A* reconciliation *R of gene tree G and species tree S is an extension of G that is a* dls*-history consistent with S.*

The parsimony criteria used to choose among the large set of possible reconciliations are usually the number of duplications (*duplication cost*), the number of losses (*loss cost*) or the sum of the two (*mutation cost*). Many algorithms have been developed for computing the most parsimonious reconciliation, the most efficient ones with running time proportional to the size of the gene tree [12, 23, 30, 67].

2.5 Perspectives on homology

There have been many uses of the word homology and the related concepts, the confusion due to the many possible measures of similarity between genes. Indeed, evolutionary, sequence, functional, or positional constraints give rise to definitions that are unfortunately not equivalent [37]. In this paper we adopt the original definitions recommended by Fitch [26], corresponding to the evolutionary concepts.

Definition 3 (homology) *Two genes are* homologous *if and only if they are the leaves of a* dls*-history H. A* gene family *is a set of homologous genes.*

Although many genes share a common origin [56], and thus share the same *dls*-history, the definition of homology given by Fitch does not include a necessary limit on the evolutionary closeness between two homologous genes. To our knowledge, this is an unfortunate and unstated ambiguity that we must live with for the time being.

The remainder of the definitions describe a hierarchy of homologous genes, implied by the *dls*-history H.

Definition 4 (orthology) *Genes a and b are* orthologous *if $lca_H(a, b)$ is a speciation node.*

As duplications may arise following a speciation event, the orthology relationship is not transitive. This property is inherent to the evolutionary definition of orthology, which is not a definition about the functional relationship between genes, nor the positional or direct descendant relationship. In this perspective, Fitch [26] introduced the following notion of *functional orthologs* or *isorthologs*, for a given function (in case of hemoglobin sequences for example, the function is the ability of being the adult transporter of oxygen).

Definition 5 (isorthology) *Two orthologous genes that have retained the same function $\mathcal{F}$ of their LCA in H are called* isorthologous *for function $\mathcal{F}$.*

Isorthology relation is transitive. Therefore it makes sense to speak of sets of isorthologs, or isorthogroups. Two genes are in the same *isorthogroup* if and only if they are isorthologous. Finally, we introduce the notion of paralogy.

Definition 6 (paralogy) *Genes a and b are* paralogous *if $lca_H(a, b)$ is a duplication node.*

Consider the histories from Figure 2a. Any two genes denoted by the same letter are homologous. The history for homologous gene family c serves as a good example. The gene from C_1 is orthologous with all occurrences of c in C_3 and C_4, while it is paralogous to the gene in C_2. Further, the last occurrences of c in C_4 is paralogous to the second occurrence of the gene in C_3.

3 Gene Order Inconsistency

In this section we explore how information on gene order can be used to discover erroneous gene trees. The general idea is the following: look at the *regions* (formally defined below) surrounding the genes of interest. If they are similar (in term of gene order), assuming that this cannot happen by chance, we can deduce that they are *homologous*, i.e. they descend, through a duplication or

speciation event, from a common ancestral region. Such property on homology for regions leads to properties on underlying genes: homologous genes in the two regions are either all pairwise orthologous or all pairwise paralogous. These properties can then be checked against gene trees, and used as criteria for correcting them.

In Section 3.1 we formally define homology on regions. This perspective allows us to establish in Sections 3.2 and 3.3 properties that sets of true gene trees must possess when genes belong to similar regions, given that the following hypothesis about convergent evolution is assumed:

Hypothesis NoConvergentEvol: Similar regions are homologous.

In the last fifteen years many methods have been developed for the classification of similar syntenic regions that have undergone gene order mutation [4, 6, 5, 34]. Hoberman and Durand [35] give a nice treatment of the competing interests surrounding a good definition of gene order similarity. David Sankoff has been ever present in the discussion [18, 36, 50, 63, 65, 66]. Whatever the definition, the underlying idea is to maximize the probability that similar regions are indeed homologous.

Our study in Section 3.4 limits regions to the immediate left and right neighbors of the genes in question; the regions of two homologous genes are similar if they are directly surrounded by homologous genes. Under this definition, the substrings *aba* of region C_5 and *aba* of region C_6 from Figure 2 are similar, as do *abc* of C_4 and *cba* of C_3.

3.1 Region homology

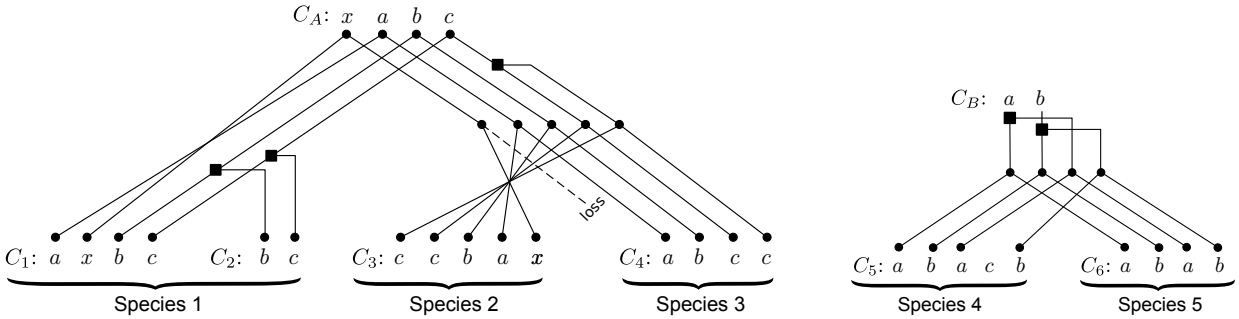
Homology on a set of genes is a property of the true history for that set, independent of any similarity measure amongst them. Homology of a set of regions should also be defined in a manner that is independent of any particular similarity measure on those regions. To accomplish this we leverage the duplication/loss/speciation histories for the genes contained in the regions of interest.

A *region* of a genome A is simply a subsequence of A. An *ancestral region* is a region occurring in some ancestral genome, while a *modern region* is a region occurring in some modern genome.

Definition 7 (region homology) *Let C_k and C_ℓ be two modern regions defined on a gene set Γ, subdivided into the gene families $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_m\}$. Let $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$ be the* dls*-histories corresponding to Γ_is, and let a_i be the root of H_i. Then C_k and C_ℓ are* homologous *if and only if the a_is all belong to a region $C_A = a_1 a_2 \cdots a_m$ of an ancestral genome A, and they are either all speciation nodes or all duplication nodes. We call C_A the LCA region for C_k and C_ℓ.*

The case where the roots of the *dls*-histories are speciations corresponds to the divergence of C_A through a speciation event, while the latter case corresponds to the divergence through a duplication event that has duplicated the entire ancestral region C_A.

Notice that the definition of region homology supports the possibility of rearrangements occurring during the evolution of regions; in Figure 2a genes a and x have been inverted in the branch from the ancestral genome to Species 1, yet regions C_1 and C_3 are homologous. Local duplications of sub-regions (in tandem or not) are also supported. In Figure 2a for example, a duplication of gene c occurs in the branch leading to Species 2 and 3, yet regions C_1 and C_3 are homologous. Insertion and deletion of genes are supported as well. For example, gene c in Species 4, which is not present in Species 5, does not prevent regions C_5 and C_6 from being homologous. Moreover,



(a) evolution following a speciation event affecting an ancestral region $C_A = xabc$.

(b) evolution following the duplication of an ancestral region $C_B = ab$.

Figure 2: Gene trees for two different ancestral regions. Duplications are denoted by square nodes, speciations by circles, and losses by dashed edges. Next to each C_i is a description of a substring of a genome. Each region C_i is defined as the genes labeling the leaves of a gene tree. For example $C_5 = abab$ (does not include gene c).

the ancestral region C_A may contain genes that have lost in the *dls*-histories leading to moderm regions.

Notice however that, in contrast to the homology relationship on genes, the homology relationship on regions is not transitive. Consequently, we are unable to generalize the notion of gene families to the notion of homologous region families.

3.2 Homology contradiction

Our definition of homologous regions, along with Hypothesis NoConvergentEvol, provides us with a tool for testing the validity of gene trees. Remember that for a pair of homologous regions, the roots of the genes trees that comprise the genes contained in the two regions must all be the same type of node; they must all be speciation nodes, or they must all be duplication nodes. Thus, for a pair of similar regions — assumed from Hypothesis NoConvergentEvol to be a pair of homologous regions — and an inferred set of gene trees — implying a set of homology relationships between genes of the regions — we can confirm that indeed the gene trees have such roots. If they do not, we say that the forest of gene trees exhibits a *homology contradiction*.

3.3 Region overlapping

In this subsection, we define the notion of a *region surrounding a gene* in a strict way ensuring a single region assignment for each gene, and a fixed length for all regions. Formally, for a given set of parameters $0 < l_1 < \cdots < l_p$ and $0 < r_1 < \cdots < r_q$, the region C_x surrounding the gene at position x in genome A is the subsequence $A[x - l_p, \cdots, x - l_1, x, x + r_1, \cdots, x + r_q]$. In Section 3.4, the underlying parameters are $p = q = 1$, and $l_1 = r_1 = 1$. Now two regions C_k and C_ℓ are *similar* if an only if, for any i, $C_k[i]$ and $C_\ell[i]$ belong to the same gene family. This definition of similarity ensures transitivity, which allows to define a *similarity family* as a family of pairwise similar regions.

In this subsection, a stronger statement on no convergent evolution is also required:

Hypothesis StrongNoConvergentEvol: Two similar regions are homologous. In addition their similarity is inherited from their LCA region and preserved during the course of evolution.

Stated formally, let C_k and C_ℓ be two similar regions surrounding two homologous genes x_k and x_ℓ belonging to a gene family Γ, and let G be the true gene tree for Γ. Then the regions surrounding ancestral genes corresponding to the nodes on the path between x_k and x_ℓ in G are similar to C_k and C_ℓ.

Take a gene tree G such that each gene (leaf of G) is assigned to a region, and that regions are grouped into similarity families $\mathcal{E} = \{F_1, F_2, \ldots, F_p\}$.

Let $V(G)$ be the set of internal nodes of G. Consider the *region labeling function* $\ell_G : V(G) \to 2^{\mathcal{E}}$ (where $2^{\mathcal{E}}$ is the power set of $\mathcal{E}$) that labels the nodes of G with homologous families as follows:

1. for all $x \in V(G)$, initialize $\ell_G(x)$ to $\emptyset$;
2. for each family F_i, include F_i in the label of any node on a path from a pair of leaves with label F_i.

The following lemma provides a second criterion for error detection in gene trees.

Lemma 1 *If G is the true gene tree for some set of genes and* Hypothesis StrongNoConvergentEvol *holds, then for each node x of G, $|\ell_G(x)| \leq 1$.*

Proof: Let x be an internal node of G with surrounding region C_x, and suppose $\ell_G(x)$ contains at least two elements F_i, F_j of $\mathcal{E}$. From the definition of ℓ_G, it follows that x is on the path between some genes ℓ_i and r_i with regions C_i^ℓ and C_i^r, both belonging to F_i. In the same manner, x is on the path between genes ℓ_j and r_j with region C_j^ℓ and C_j^r belonging to F_j. We have that x has at least one descendant that is ℓ_i or r_i, and at least another descendant that is ℓ_j or r_j. Suppose without loss of generality that ℓ_i and ℓ_j are descendants of x. By Hypothesis StrongNoConvergentEvol, C_i^ℓ and C_j^ℓ are both similar to C_x, and since similarity is transitive, C_i^ℓ is similar to C_j^ℓ. It follows that $F_i = F_j$. □

A gene tree with an internal node possessing multiple labels is said to exhibit a *region overlap*. Notice that for such a node, Lemma 1 holds whether it is a speciation or a duplication. Figure 3 shows a gene tree with multiple region overlaps, which are all duplications. Consider the overlapping occurring at the root of G, which we denote by r. It might be tempting to explain this scenario by stating that since r is a duplication, one copy of the ancestral gene belonged to the ancestral region similar to F_1, and the other to the ancestral region similar to F_2, and thus both regions could have propagated to their respective descendants. However, r refers to a single ancestral gene, which may have belonged to one of the two ancestral regions, but not to both, as we assume each gene is assigned a single region.

3.4 Results

We wanted to see the impact of using homology contradiction and Lemma 1 to reveal errors in gene trees. To this end, we considered the four fish genomes *Gasterosteus aculeatus* (Stickleback), *Oryzias latipes* (Medaka), *Tetraodon nigroviridis*, and *Danio rerio* (Zebrafish) with human and

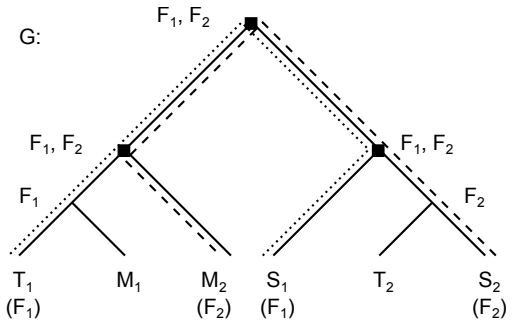


Figure 3: The gene tree of the "RAB27" gene family (ENSGACG00000003336) for the Stickleback (S), Medaka (M) and Tetraodon (T) species, exhibiting a region overlapping. The T_1, S_1 genes are in similarity family F_1, while M_2, S_2 are in another similarity family F_2. The internal nodes are annotated by their l_G labeling; all nodes on the dotted path are labeled by F_1, and those on the dashed path by F_2.

mouse as outgroups. We used the *Ensembl Genome Browser* to collect all available gene trees, and filtered each tree to preserve only genes from the taxa of interest. We then reconciled the trees with the known species trees, and identified duplication and speciation nodes. Genes appearing in the same gene tree in the database are considered to be part of the same homologous gene family.

In this section, a region surrounding a gene is defined as the substring containing the gene and both its left and right adjacencies. Two regions are similar if they contain homologous genes in the same order or inverted order. More precisely, regions $C_k = x_1a_1y_1$ and $C_\ell = x_2a_2y_2$ (or $C_\ell = y_2a_2x_2$) are similar if x_1 and x_2 appear in the same *Ensembl* gene tree, a_1 and a_2 appear in the same gene tree, and y_1 and y_2 appear in the same gene tree. We avoid tandem duplications by requiring the three trees to be different.

In Section 3.2 we defined the homology contradiction property for a forest of gene trees. Here, we identify problematic forests of gene trees using that property. Let $C_k = x_1a_1y_1$ and $C_\ell = x_2a_2y_2$ be two similar regions and G_x, G_a, and G_y be the gene trees containing the pairs of homologs (x_1, x_2), (a_1, a_2) and (y_1, y_2) respectively. Then, according to our definition, the forest $\{G_x, G_a, G_y\}$ exhibits a homology contradiction iff the set $\{lca_{G_x}(x_1, x_2), lca_{G_a}(a_1, a_2), lca_{G_y}(y_1, y_2)\}$ contains at least one duplication node and at least one speciation node.

In this section we will focus on the gene tree of the central gene. We say that G_a exhibits a *paralogy contradiction* iff $lca_{G_a}(a_1, a_2)$ is a duplication node, and both $lca_{G_x}(x_1, x_2)$ and $lca_{G_y}(y_1, y_2)$ are speciation nodes. Conversely, we say that G_a exhibits an *orthology contradiction* iff $lca_{G_a}(a_1, a_2)$ is a speciation node, and both $lca_{G_x}(x_1, x_2)$ and $lca_{G_y}(y_1, y_2)$ are duplication nodes. Note that this notion of contradiction is extremely conservative; if only a single neighbor disagrees with the central gene, then we do not report it.

Results are summarized in Table 1. Among the 6241 trees in Ensembl, 6118 of them have at least one pair of genes in the same context. More than 31% of the 6241 trees exhibited at least one contradiction, the most frequent contradiction type being paralogy contradiction. These numbers show that a very conservative application of our methods uncovers a significant number of inconsistencies between gene order and gene tree topology.

It is conceivable that a significant number of missing genes in the gene trees could lead to a

Number of trees	6241
Region overlap	3.4% (210)
Paralogy contradiction	22.5% (1407)
Orthology contradiction	10.8% (677)
At least one contradiction	31.3% (1959)

Table 1: Results obtained for Ensembl gene trees. Reported numbers are not mutually exclusive, in the sense that a given tree may exhibit more than one type of contradiction, and thus be included in more than one list. In brackets are the actual numbers of trees.

false homology contradiction. Also, poor detection of homology relationships in Ensembl could yield false region overlaps. For example, two overlapping regions could have the form $C_k = a_1 b_1 c_1$ and $C_\ell = x_2 b_2 c_2$. But if x_2 should in fact be in the same homologous gene family as a_1, the overlap would no longer exist. This is what happens in the example of Figure 3. The F_1 region consists of "ASH1L" "RPS27" "KCNN3" genes, while the F_2 region is made of "RAB13" "RPS27" "KCNN3" genes. In fact, every single overlapping regions we found had this form. Thus region overlaps in Ensembl gene trees might not occur because of wrong topologies, but rather because of missing homologies. In any case, detection of overlaps can identify possible improvements on the known relationship between some pairs of genes.

To get an idea of how the numbers can change, we reran the test suite for a more general notion of similarity: C_k and C_ℓ are similar if b_1 and b_2 are homologous, and if there exists a pair of neighbors c_1 and c_2 that are homologous. Note that under this definition, there are fewer region sets so region overlaps are harder to find. The new definition finds 71 (2.38%) gene trees with overlaps.

Yet our region overlaps and homology contradictions tend to agree with mechanisms already in place for error detection in Ensembl gene trees. Based on the structure of the tree, some duplication nodes, corresponding to NAD nodes [12], are labeled as "dubious" in the Ensembl trees. As paralogy and orthology contradictions are inferred according to duplication nodes (one duplication node involved in a paralogy contradiction and two in an orthology contradiction), we were interested to see to which extent our results correlated with Ensembl observations about dubious duplications. We found that 77.4% of duplications involved in observed paralogy contradictions are labeled as dubious, while 90.2% of duplications involved in orthology contradictions are dubious. These number are significantly high considering that the fraction of dubious duplications among the total number of duplications in our trees is only 36%. These observations validate the fact that gene order inconsistencies are likely to reveal errors in gene trees.

4 Positive and Negative selection Bias

Classical phylogenetic methods, such as those using parsimony, distance or maximum likelihood models, are typically based upon the assumption of stochastic, neutral, and site-independent processes. However, as few mutations may cause structural modification to protein coding genes with deleterious functional consequences, isorthologous gene copies in multiple species are commonly subject to negative (purifying) selection pressure, leading to sequence stability inside isorthogroups.

On the other hand, positive selection, responsible for the creation of new function, is also known to play a major role in the evolution of gene families. Under natural (positive and negative) selection, a gene tree best reflecting the sequence similarity of gene copies is more likely to reflect functional constraints rather than evolutionary and ancestral relationships between gene copies. In particular, negative selection may result in isorthologous genes being grouped into a subtree of the gene tree, leading to erroneous ancestral inference for the isorthogroup.

This grouping driven by function has been reported for different gene families, such as GLP-1 [51] and opsin proteins [55]. An interesting study based on simulations is also reported in [40]. In this study, DNA sequences encoding a protein folding, with a predefined active site for the binding of a ligand, have been generated. An A ligand initially bound stably at the beginning of the simulation, while a B ligand did not. The proteins were evolved under constant population size and mutation rate. In every generation the individuals were picked randomly, provided they folded stably and binded to a peptide. Moreover, to simulate positive selection, a selective advantage of 5% was given to individuals binding the new ligand B. Phylogenetic trees for simulated sequences were then inferred using distance, parsimony and likelihood methods. Every generated tree exhibited a clustering by function rather than by ancestry (two monophyletic groups, one for proteins binding to the ligand A and the other for proteins binding to the ligand B). In the same paper, other results obtained on multiple sequence alignments of Chordate genes also confirmed previous studies on the loss of the evolutionary signal due to negative and positive selection [48, 58, 32].

4.1 Detecting functional bias

In the presence of negative and positive selection (*i.e.* confusion of the neutral phylogenetic signal), some studies have recommended different criteria for gene (site) selection when reconstructing phylogenies. In particular, the filtering of fast evolving genes has been suggested to reduce the effect of positive selection [32]. On the other hand, filtering slow evolving sites has been suggested to reduce the effect of negative selection. However, as noticed in [40], these models for data filtering have limitations as evolution speed does not always correlate with selection type.

Instead of an *a priori* selection of appropriate sites, we can alternatively *a posteriori* detect gene trees reflecting a bias due to negative or positive selection. Classical methods for evaluating selective pressures acting on homologous amino acid sequences are based on computing the ratio dN/dS of the number of non-synonymous (dN) versus synonymous (dS) nucleotide substitutions per site of a pairwise alignment [41]. Synonymous substitutions are those that do not result in change of amino acid (for instance most changes at the third codon position), while non-synonymous substitutions are those altering the amino acid (for instance changes at the second codon position). Under negative (purifying) selection, most non-synonymous changes are eliminated, leading to an excess of synonymous changes. On the other hand, positive selection leads to an excess of non-synonymous substitutions. In general, negative selection is inferred if $dN/dS < 1$ and positive selection is inferred if $dN/dS > 1$. We suggest the use of the synonymous/non-synonymous substitution rate measure for detecting gene trees reflecting a selection bias, formalized as trees reflecting the *isolocalization property* which is defined below.

4.2 Formalizing the functional bias

Under the hypothesis that after a duplication, exactly one of the two gene copies preserve the parental function, the *isolocalization property* was introduced in [54], to characterize gene trees biased towards a grouping of isorthologous genes. Here, we define a less constraining version of this property by asking for at least one isorthogroup to appear as a monophyletic group (an isolated subtree). Notice that results obtained in [54] (stated below and summarized in Section 5) about the effect on reconciliation remain valid for this new definition.

Definition 8 (isolocalization) *Let G be a gene tree for a gene family Γ. Let $I = \{a_1, a_2, \ldots, a_n\} \subseteq \Gamma$ be a maximal isorthogroup of Γ, meaning that no other gene of Γ is isorthologous to an a_i. A gene tree G respects the* isolocalization *property for I if and only if there exists an x such that $\mathcal{L}(G_x) = I$.*

We say that G respects the isolocalization property *if G respects the isolocalization property for at least one maximal isorthogroup of Γ.*

We showed in [54] that isolocalization confounds reconciliation, in the sense that some histories (those with a duplication node descending from a speciation node) can never be recovered through the reconciliation of a gene tree respecting the isolocalization property. Following this observation, we proposed general ideas for inferring true histories. Although presented as tools for correcting reconciliation, they can alternatively be seen as tools for correcting gene trees, *i.e.* removing the functional contraints exhibited by isorthogroups. An overview of the related open problems is given in Section 5.

In the following, an *isorthologous subtree* of G is a speciation subtree of G with a set of leaves corresponding a maximal isorthogroup.

4.3 Results

By definition, a subtree G_x rooted at node x of a gene tree G is an *isorthologous subtree* if $\mathcal{L}(G_x)$ is a maximal isorthogroup, *i.e.* elements of $\mathcal{L}(G_x)$ are pairwise isorthologous, and there is no gene outside $\mathcal{L}(G_x)$ which is isorthologous to a gene of $\mathcal{L}(G_x)$. As suggested by the discussion above, this can be tested by comparing the dN/dS ratios of pairs (I_i, I_j) of genes inside $\mathcal{L}(G_x)$, versus pairs (I_k, O_l), with I_k being a gene inside $\mathcal{L}(G_x)$ and O_l being a gene outside $\mathcal{L}(G_x)$. Here, we consider the average dN/dS ratios over all possible pairs. Namely, we define M_x^I to be the average over all (I_i, I_j) *inside pairs* and M_x^O to be the average over all (I_k, O_l) *inside-outside pairs*. For an isorthologous subtree, we expect $\frac{M_x^I}{M_x^O}$ to be lower than one. For any internal node x, if $\frac{M_x^I}{M_x^O} < 1$ we say x is a *winner*; otherwise we say that x is a *loser*. Note that the root of a tree cannot be a winner, since there are no genes outside of its leafset.

We wanted to see to what extent the Ensembl gene trees reflect a natural selection bias. We considered the same six species as in Section 3.4, namely four fish species (Stickleback, Medaka, Tetraodon, Zebrafish) with human and mouse as outgroups. We collected all available gene trees, restricted each of them to the taxa of interest, reconciled the trees with the known species trees, and retained the "interesting" ones according to [54], namely those reflecting a history with a "surviving" duplication followed by a "surviving" speciation event. More precisely, a gene tree G was retained if it contained at least one duplication node x such that G_{x_ℓ} and G_{x_r} were both speciation subtrees, each containing at least two leaves and at least five leaves together. This

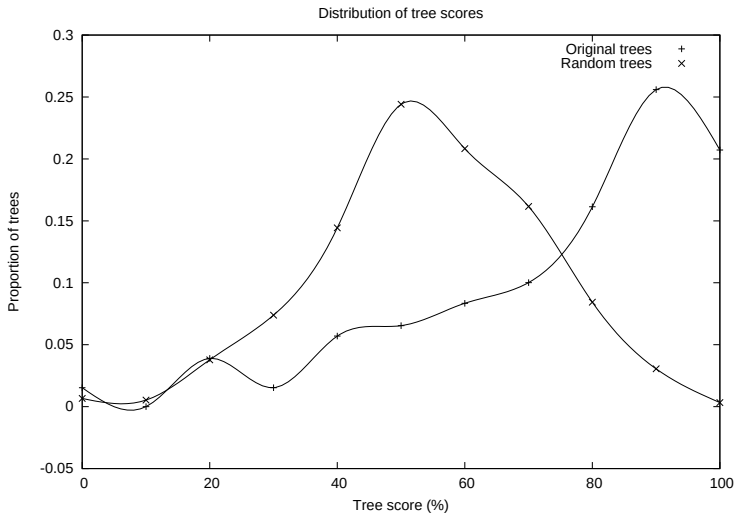


Figure 4: Distribution of original trees scores versus random trees scores. The score of a tree is its number of winner nodes over its number of internal nodes.

yielded 815 gene trees. We refer to this set as the *original set*. For each tree G in the original set, we obtained the canonical nucleotide sequences of its genes from Ensembl, and computed every pairwise dN/dS ratio using the PAML package [64], which implements the Nei and Gojobori method [42]. The sequences were aligned and prepared using ClustalW2 [20] in conjunction with the BioPerl library [19].

We expect the topology of each tree G in the original set to contain more winner nodes than most other topologies that share the same leaf set. We tested the null hypothesis, which states that there is no relationship between the gene trees constructed by Ensembl and the proportion of winner nodes they contain. Thus for each tree G, we considered a set of *random trees*, obtained from G by all possible permutations on its leaves. We refer to the set of random trees for all the Ensembl trees as the *random set*.

Figure 4 depicts, for both the original and random tree sets, the proportion of trees by *score*, defined for each tree as the number of winners over the number of internal nodes. The original trees clearly tend to contain a higher ratio of winners than random trees. In fact, the random trees' percentages follow a distribution that is not far from normal, whereas the original trees favor higher scores, hinting at the invalidity of the null hypothesis.

Our analysis also showed some interesting numbers. Among all nodes (excluding roots and leaves), 71% of them are winners in the original set, as compared to 50% in the random set. Moreover, 81% of the original trees have a majority of winner nodes (more than half), compared to 49% for random trees. Say that a gene tree G is *optimal* if the number of winner nodes in G is no less than the number in all random trees for G. We find that the proportion of optimal gene trees over the original set is 45%. Moreover, if we also count as winner nodes those having a winner ancestor (*i.e.* not only those pointing to isorthogroup but also to subsets of isorthogroups), then the proportion of optimal trees raises to 64% of all original trees. Finally, 80% of the original trees have more winner nodes than at least half of their random trees.

More detailed statistical analysis are required to establish criteria for detecting functional bias in a gene tree according to dN/dS ratios. However, this preliminary study already reveals a possible negative selection bias in these Ensembl trees.

5 Gene Tree Correction

A significant obstacle to our understanding of evolution is the difficulty of inferring accurate gene trees. It is now clear that methodology based solely on sequence similarity are unable to produce a single well supported gene tree [44, 45, 59, 61]. Opposite to such a "sequence only" paradigm is the "sequence free" paradigm that does not directly use the sequence information. An example is the polynomial-time algorithm developed by Durand *et al.* [17] for inferring a gene tree minimizing the reconciliation cost with a given species tree. Such an extreme strategy is of theoretical interest only, as an accurate reconstruction model should be "hybrid", e.g. account for both sequence and genomic information, the challenge being to find the right balance between the two. Later in the same paper, a hybrid approach is in fact presented.

Each one of the genomic constraints we have introduced in this paper can be used to define, in the space of gene trees, points that best reflect the desired properties. As exploring the space of all topologies is time and space prohibitive, gene tree correction methods explore the neighborhood of an input gene tree G, according to a tree-distance measure, such as the Robinson-Foulds [11, 29], Nearest Neighbor Interchange (NNI) [13, 28, 27], Subtree Prune and Regraft (SPR), or Tree Bisection and Reconnection (TBR) [10] distances. In order to reduce the space of explored gene trees, tree moves may be restricted to edges deemed suspect by the user, typically those with low bootstrap values [13, 17].

As in Durand *et al.*, almost all hybrid methods that have been developed so far are "species tree-aware" and consist in selecting, from a given neighborhood, a tree minimizing a reconciliation distance with a species tree. Beside reconciliation, other criteria such as the number of NAD nodes [12, 16, 53] may be considered for a "species tree-aware" hybrid method. On the other hand, a "gene order aware" method would select, in a given neighborhood of G, the trees avoiding or minimizing gene order inconsistencies (Section 3). A "negative selection aware" method would select appropriate alternative trees, as we explain in Section 5.

A wide range of theoretical and analytical open problems are implicit in the last paragraph. In addition to developing the right data structures and algorithms for efficient exploration of the neighborhood of a gene tree, the challenge is to explore ways of combining multiple criteria in a unified framework. Do repairs to a gene tree suggested by the diversity of constraints coincide, or do they conflict? If they conflict, how should relative importance be distributed over the various constraints?

Another concern is the development of a unified approach that accounts for both sequence and genomic constraints simultaneously. Indeed, a significant drawback of the hybrid methods developed so far is the sequential manner in which the sequence and genomic information are considered; the corrected gene tree is not subsequently evaluated according to the sequence information, and thus may over fit the species tree. From this perspective, an interesting framework is the one used in TreeFix [62], as well as PhylDog [8] and Spimap [45]. Taking advantage of the fact that phylogenetic methods usually lead to a set of statistically equivalent gene trees, TreeFix is based on a heuristic that searches, among all topologies that are statistically equivalent to the input tree,

one that minimizes a user-defined reconciliation cost. The implicit hypothesis used in TreeFix is that regions of tree space with high sequence likelihood and low reconciliation cost overlap, which they show to be true in practice. Such a general framework can easily be adapted to account for various types of constraints. However, the more constraints simultaneously considered, the more challenging the problem of attributing relative weights to each of them and managing conflicting requirements become (see also chapter Chauve *et al.* in this volume).

We conclude this section by highlighting important results obtained in [54] that show how the selection bias, formalized as the isolocalization property, can be used for gene tree correction.

Isorthology respecting histories

As recalled in Section 4.2, we showed that gene trees respecting the isolocalization property can lead to erroneous histories through reconciliation. This observation is not surprising as a gene tree reflecting functional constraints rather than evolutionary constraints can hardly be confidently used to infer evolutionary scenarios. Yet there must be some information in the gene tree and species tree relationship. For instance, we expect subtrees corresponding to isorthogroups in a well-supported gene tree to agree with the species tree. Define a *speciation subtree* of G to be a subtree such that all internal nodes (if any) are labeled as speciations by the reconciliation. The following result comes from Corollary 3 of [54], and is adapted to our new definition of the isolocalization property.

Theorem 1 *Let G be a gene tree satisfying the isolocalization property for an isorthogroup I and reflecting the true phylogeny for I (see a precise definition in [54]). Then I appears in G as the leaf-set of a speciation subtree.*

Based on Theorem 1, the following definition can be used for gene tree correction.

Definition 9 (isorthology respecting history (IRH)) *Given a gene tree G and a species tree S, a* dls*-history H is an* isorthology respecting history *for (G, S) if and only if each isorthogroup inferred from H is the leaf-set of a speciation subtree of G.*

Following a duplication, we assume that one of the two gene copies preserves the ancestral function (Hypothesis 1 in [54]). Suppose that gene related by speciation preserve the ancestral function. Then two isorthogroups $\{M1, S1, T1, Z1\}$ and $\{M2, S2\}$ are inferred from the history H in Figure 5, and H is an isorthology respecting history for (G, S). Notice that H leads to the gene tree G', which can be seen as a correction of G.

As many IRHs are possible for a given pair (G, S), an appropriate criterion for choosing most likely histories is required. For example the history R resulting from the reconciliation of G with S in Figure 5 is also an isorthology respecting history for (G, S). However, while R has a mutation cost of 3 (one duplication and two losses), the history H has a mutation cost of one (no loss). In [54] we considered the Minimum Isorthology Respecting History Reconstruction (MIRH) Problem, which asks for the IRH of minimum cost, and developed a linear-time algorithm for the duplication cost. An algorithm for the mutation cost remains open.

The MIRH optimization problem as stated, is very conservative, in the sense that nothing is trusted in the gene tree except the isorthology information. In particular, it ignores all the information on duplication and speciation nodes of G that are above the considered speciation subtrees. An alternative would be to account for the hierarchy of deeper nodes in G. The notion

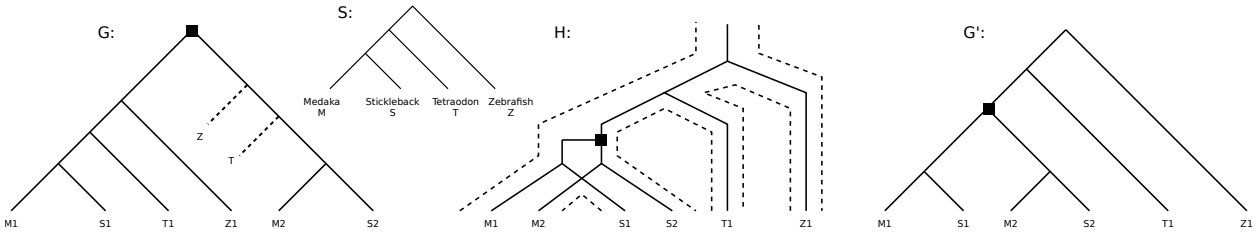


Figure 5: G is the gene tree for the gene family "C2orf47" from Ensembl (ENSGT00390000004145), extended with loss leaves according to a reconciliation with species tree S. M is Medaka, S is Stickleback, T is Tetraodon, and Z is Zebrafish. Reconciliation of G with the species tree S gives one duplication and two losses (as marked in G, duplication by a square, and losses by dotted lines). Considering the largest speciation subtrees of G as pointing to the isorthogroups ($\{M1, S1, T1, Z1\}$ and $\{M2, S2\}$), H is an isorthology respecting history for (G, S) leading to the gene tree G'.

of a *Triplet Respecting History* (TRH) [54] is intended to account for such hierarchy. Efficient algorithms for inferring parsimonious TRHs remain undiscovered.

Notice that Theorem 1 does not *a priori* give us the isorthogroups for a pair (G, S), as the true isorthologous subtree could be part of a larger speciation subtree. A restricted version of the MIRH problem considers the maximal speciation subtrees of G as the definition of the isorthogroups. We showed in [54] that this isorthology respecting partition of G is the one that would minimize the duplication cost, but not necessarily the mutation cost.

An alternative approach would use some isorthogroup detection criteria, such as the one given in Section 4.1, and correct according to the corresponding isorthologous subtrees. Such targeted reconstruction algorithms remain completely unexplored.

6 Conclusion

While gene trees have traditionally been constructed and validated using nucleotide sequence or amino acid sequence information alone, more recently information from the species tree has been used to both correct and validate gene trees. We have introduced new methodology to further validate and correct gene trees through the use of other data. Our novel use of syntenic information (homologous regions) points to a significant number of flawed gene trees in the Ensembl database due to homology contradiction or region overlapping. Our use of the dN/dS ratio on gene trees points to a bias towards clustering of isorthologous genes in gene trees. Although some potential avenues for improving gene trees are explored, our results seem to pose more questions than they answer.

References

[1] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Simultaneous bayesian gene tree recons. and reconciliation analysis. *Proc. Nat. Acad. Sci.*, 106(14):5714-5719, 2009.

[2] L. Arvestad, A.C. Berglund, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB*, pages 326-335, 2004.

[3] R.G. Beiko and N. Hamilton. Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.*, 6(15), 2006.

[4] A. Bergeron, C. Chauve, and Y. Gingras. Formal models of gene clusters. In I. Mandoiu and A. Zelikovsky, editors, *Bioinformatics algorithms: techniques and applications*, chapter 8. Wiley, 2008.

[5] A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. *Algorithms in Bioinformatics*, pages 464–476, 2002.

[6] A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *Journal of Computational Biology*, 13:1340–1354, 2003.

[7] A.C. Berglund-Sonnhammer, P. Steffansson, M.J. Betts, and D.A. Liberles. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *Journal of Molecular Evolution*, 63:240-250, 2006.

[8] B. Boussau, G.J. Szollosi, and L. Duret *et al.* Genome-scale coestimation of species and gene trees. *Genome Research*, 23:323-330, 2013.

[9] W.C. Chang and O. Eulenstein. Reconciling gene trees with apparent polytomies. In D.Z. Chen and D. T. Lee, editors, *Proceedings of the 12th Conference on Computing and Combinatorics (COCOON)*, volume 4112 of *Lecture Notes in Computer Science*, pages 235–244, 2006.

[10] R. Chaudhary, J.G. Burleigh, and O. Eulenstein. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC-Bioinformatics*, 13(Supp.10):S11, 2011.

[11] R. Chaudhary, J.G. Burleigh, and D. Fernandez-Baca. Fast local search for unrooted Robinson-Foulds supertrees. *IEEE/ACM Transactions on Comutational Biology and Bioinformatics*, 9(4):1004-1012, 2012.

[12] C. Chauve and N. El-Mabrouk. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. In *RECOMB 2009*, volume 5541 of *LNCS*, pages 46-58. Springer, 2009.

[13] K. Chen, D. Durand, and M. Farach-Colton. Notung: Dating gene duplications using gene family trees. *J. Comp. Biol.*, 7:429–447, 2000.

[14] R. Dondi and N. El-Mabrouk. Minimum leaf removal for reconciliation: Complexity and algorithms. In *CPM*, volume 7354 of *Lecture Notes in Computer Science*, pages 399-412. Springer, 2012.

[15] R. Dondi, N. El-Mabrouk, and K.M. Swenson. Gene tree correction for reconciliation and species tree inference: complexity and algorithms. *Journal of Discrete Algorithms*, (submitted), 2013.

[16] A. Doroftei and N. El-Mabrouk. Removing noise from gene trees. In *WABI*, volume 6833 of *LNBI/LNBI*, pages 76-91, 2011.

[17] D. Durand, B.V. Haldórsson, and B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13:320–335, 2006.

[18] D. Durand and D. Sankoff. Tests for gene clustering. *Journal of Computational Biology*, 10(3-4):453–482, 2003.

[19] J.E. Stajich *et al.* The bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12:1611- 1619, 2002.

[20] M.A. Larkin *et al.* Clustalw and clustalx version 2. *Bioinformatics*, 23:2947- 2948, 2007.

[21] P. Flicek *et al.* Ensembl 2012. *Nucleic Acids Research 2012 40 Database issue*, 40:D84- D90, 2012.

[22] D.L. Swofford *et al.* PAUP: phylogenetic analysis using parsimony. 4th ed., Sinauer Associates, 2002.

[23] O. Eulenstein, B. Mirkin, and M. Vingron. Duplication-based measures of difference between gene and species trees. *Journal of Computational Biology*, 5:135– 148, 1998.

[24] J. Felsenstein. Evolurionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368-376, 1981.

[25] J. Felsenstein. PHYLIP(phylogeny inference package). Version 3.6. distributed by the author, Seattle (WA): Department of Genome Sciences, University of Washington, 2005.

[26] W. M. Fitch. Homology. a personal view on some of the problems. *TIG*, 16(5):227- 231, 2000.

[27] P. Gorecki and O. Eulenstein. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics*, 13(Supp 10):S14, 2011.

[28] P. Gorecki and O. Eulenstein. A linear-time algorithm for error-corrected reconciliation of unrooted gene trees. In *ISBRA*, volume 6674 of *LNBI*, pages 148-159. Springer-Verlag, 2011.

[29] P. Gorecki and O. Eulenstein. A Robinson-Foulds measure to compare unrooted trees with rooted trees. In L. Bleris *et al*, editor, *ISBRA*, volume 7292 of *LNBI*, pages 115-126, 2012.

[30] P. Gorecki and J. Tiuryn. DLS-trees: a model of evolutionary scenarios. *Theoretical Computer Science*, 359:378–399, 2006.

[31] S. Guidon and O. Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696- 704, 2003.

[32] H. Philippe H, P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. Early-branching or fast-evolving eukaryotes? an answer based on slowly evolving positions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267:1213- 1221, 2000.

[33] M.W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology*, 8(R141), 2007.

[34] S. Heber and J. Stoye. Algorithms for finding gene clusters. *Algorithms in Bioinformatics*, pages 252–263, 2001.

[35] R. Hoberman and D. Durand. The incompatible desiderata of gene cluster properties. *Comparative Genomics*, pages 73–87, 2005.

[36] R. Hoberman, D. Sankoff, and D. Durand. The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology*, 12(8):1083–1102, 2005.

[37] E.V. Koonin. Orthologs, paralogs and evolutionary genomics. *Annual Reviews on Genetics*, 39:309- 338, 2005.

[38] M. Lafond, K.M. Swenson, and N. El-Mabrouk. An optimal reconciliation algorithm for gene trees with polytomies. In *LNCS*, volume 7534 of *WABI*, pages 106-122, 2012.

[39] H. Li, A. Coghlan, J. Ruan, and L.J. Coin *et al.*. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 34:D572- 580, 2006.

[40] S.E. Massey, A. Churbanov, S. Rastogi, and D.A. Liberles. Characterizing positive and negative selection and their phylogenetic effects. *Gene*, 418:22- 26, 2008.

[41] T. Miyata and T. Yasunaga. Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1):23–36, 1980.

[42] M. Nei and T. Gojobori. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3:418- 426, 1986.

[43] T-H. Nguyen, V. Ranwez, S. Pointet, A-M A. Chifolleau, J-P. Doyon, and V. Berry. Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for Molecular Biology*, 8(12), 2013.

[44] M.D. Rasmussen and M. Kellis. Accurate gene-tree reconstruction by learning gene and species-specific substitution rates accross multiple complete geneomes. *Genome Research*, 17:1932-1942, 2007.

[45] M.D. Rasmussen and M. Kellis. A bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution*, 28(1):273- 290, 2011.

[46] F. Ronquist and J.P. Huelsenbeck. MrBayes3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572- 1574, 2003.

[47] R. Ruan, H. Li, and Z. Chen *et al.*. TreeFam: 2008 update. *Nucleic Acids Research*, 36(suppl. 1):D735-D740, 2008.

[48] V. Ruano-Rubio and V. Fares. Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad and the ugly. *Systematic Biology*, 56:68- 82, 2007.

[49] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425, 1987.

[50] D. Sankoff, V. Ferretti, and J. H. Nadeau. Conserved segment identification. *Journal of Computational Biology*, 4(4):559–565, 1997.

[51] M. Skovgaard, J.T. Kodra, D.X. Gram, S.M. Knudsen, D. Madsen, and D.A. Liberles. Using evolutionary information and ancestral sequences to understand the sequence-function relationship in GLP-1 agonists. *Journal of Molecular Biology*, 363:977- 988, 2006.

[52] A. Stamatakis. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. *Bioinformatics*, 22:2688-2690, 2006.

[53] K. M. Swenson, A. Doroftei, and N. El-Mabrouk. Gene tree correction for reconciliation and species tree inference. *Algorithms for Molecular Biology*, 7(31), 2012.

[54] K. M. Swenson and N. El-Mabrouk. Gene trees and species trees: Irreconcilable differences. *BMC Bioinformatics*, 13((Suppl 19)):S15, 2012.

[55] S.D. Taylor, K.D. de la Cruz, M.L. Porter, and M.F. Whiting. Characterization of the long-wavelength opsin from Mecoptera and Siphonaptera: does a flea see? *Molecular Biology and Evolution*, 22:1165- 1174, 2005.

[56] D.L. Theobald. A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–222, 2010.

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

[58] J.P. Townsend. Profiling phylogenetic informativeness. *Systematic Biology*, 56:222- 231, 2007.

[59] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara gene trees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19:327-335, 2009.

[60] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13):i549–i558, 2007.

[61] K.M. Wong, M.A. Suchard, and J.P. Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319:473-476, 2008.

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

[63] X. Xu and D. Sankoff. Tests for gene clusters satisfying the generalized adjacency criterion. *Advances in Bioinformatics and Computational Biology*, pages 152–160, 2008.

[64] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24:1586- 1591, 2007.

[65] Z. Yang and D. Sankoff. Natural parameter values for generalized gene adjacency. *Journal of Computational Biology*, 17(9):1113–1128, 2010.

[66] Q. Zhu, Z. Adam, V. Choi, and D. Sankoff. Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(2):213–220, 2009.

[67] C. M. Zmasek and S. R. Eddy. A simple algorithm to infer gene duplication and speciiation events on a gene tree. *Bioinformatics*, 17:821– 828, 2001.