Notung: Dating Gene Duplications using Gene Family Trees *

Kevin Chen,† Dannie Durand‡ Martin Farach-Colton§

Abstract

Large scale gene duplication is a major force driving the evolution of genetic functional innovation. Whole genome duplications are widely believed to have played an important role in the evolution of the maize, yeast and vertebrate genomes. The use of evolutionary trees to analyze the history of gene duplication and estimate duplication times provides a powerful tool for studying this process. Many studies in the molecular evolution literature have used this approach on small data sets, using analyses performed by hand. The rapid growth of genetic sequence data will soon allow similar studies on a genomic scale but such studies will be limited unless the analysis can be automated. Even existing data sets admit alternative hypotheses that would be too tedious to consider without automation.

In this paper, we describe a toolbox called NOTUNG that facilitates large scale analysis, using both rooted and unrooted trees. When tested on trees analyzed in the literature, NOTUNG consistently yielded results that agree with the assessments in the original publications. Thus, NOTUNG provides a basic building block for inferring duplication dates from gene trees automatically and can also be used as an exploratory analysis tool for evaluating alternative hypotheses.

1 Introduction

Yeast is a single cell organism with 6000 genes [5], while mice have an estimated 50,000 - 100,000 genes [24]. How did this order-of-magnitude increase in gene number, with its concomitant increase in functional complexity, arise? Gene duplication followed by mutation leads to new function and is considered the principal force driving developmental innovation in vertebrates [18].

The availability of sequence data has catalyzed the study of the impact of duplication, especially whole genome duplication, on the evolution of genomic structure (see [25] for a survey), as well as the specialization of function through the evolution of gene families. An important tool in the study of both questions is the construction and analysis of trees based on the sequences of duplicated genes, so called gene family trees.

Until recently, such studies involved a small number of gene families, each represented by ten or twenty sequences, and the analysis could be carried out by visual inspection of the trees [2, 8, 9, 11, 22, 23]. However, as genomic sequence data grows, the number of gene families to be considered in a single genome will grow, and so will the number of trees to be analyzed. For example, in their analysis of duplications in the yeast genome [28], Wolfe and Shields identified 446 duplicated genes. This data set is an order of magnitude larger than the gene duplication studies currently being carried out by hand.

In this paper, we formalize the analytic methods described verbally in the molecular evolution studies and cast them into a unified framework. Using this framework, we develop computational methods for analyzing duplication histories and determining duplication dates in rooted trees, as well as exploring two kinds of alternative hypotheses: alternate rootings for unrooted trees and local rearrangements when the evidence supporting an edge is weak. These methods were implemented in a set of tools called NOTUNG that can be used for exploring alternative hypotheses about duplication events and is a step towards the automated analysis of duplications in large genomic data sets.

[*] To appear in RECOMB2000, The Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan,April 8 - 11, 2000

[†] Department of Computer Science, Princeton University, Princeton, NJ 08544, USA (*kcchen@princeton.edu*).

[‡] Contact author: Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA (*durand@cs.princeton.edu, http://www.cs.princeton.edu/∼durand*).

[§] Department of Computer Science, Rutgers University, Piscataway, NJ 08855, USA (*farach@cs.rutgers.edu, http://www.cs.rutgers.edu/∼farach*).

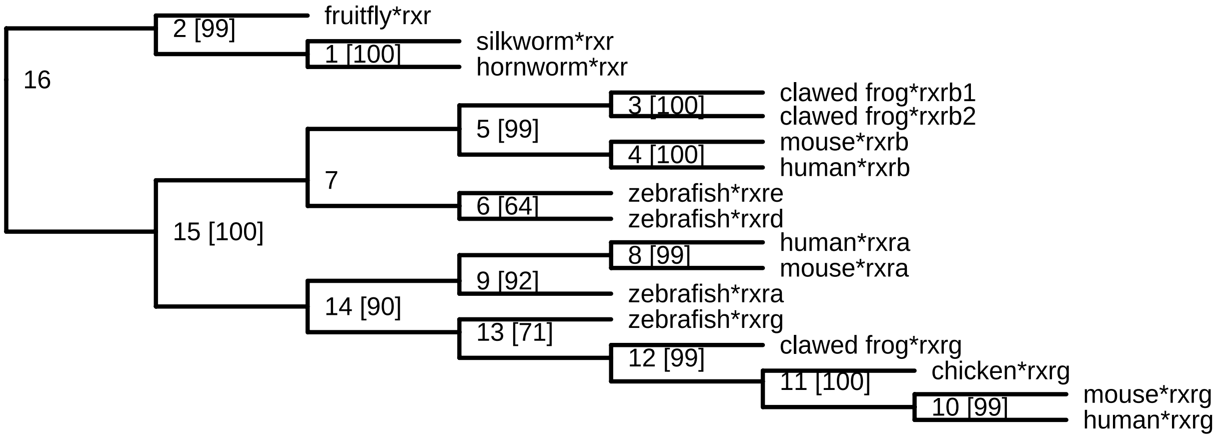

Figure 1: A rooted Neighbor Joining tree for the RXR family reproduced from [8]. Branch labels [in square brackets] represent the percentage of bootstrap samples supporting that branch. Values ≤ 50% are not shown.

An Example of Duplication Analysis. A gene family is "a set of genes descended by duplication and variation from some ancestral gene" [12], typically exhibiting related sequence and function. A *gene family tree (GFT)* is a phylogeny constructed from the sequences of family members, including representatives of the same gene in different species (orthologs) and duplicate genes in the same species (paralogs). A GFT differs from a species tree in that a species may appear more than once.

We begin by considering a typical analysis of gene duplication using a gene family tree. Hughes analyzed the evolution of the RXR family [8], using the rooted tree reproduced in Figure 1, which was constructed using the Neighbor Joining heuristic. Confidence in clustering patterns was assessed using bootstrapping, a statistical resampling method [1].

Summarizing the history of the RXR family that can be inferred from the tree, Hughes states "RXR genes from three insects fell outside of all the vertebrate RXRA, RXRB and RXRG genes. The phylogeny suggests that RXRB diverged first followed by RXRA and RXRG. ... Zebrafish genes were found to cluster with mammalian RXRB, RXRA and RXRG, but bootstrap support for these clustering patterns was not strong. Frog RXRB and RXRG genes cluster with their mammalian counterparts and, in each of these cases, there is strong (99%) bootstrap support. The tree thus suggests that RXRA, RXRB and RXRG diverged before the divergence of amphibians and amniotes and probably before the divergence of tetrapods and bony fishes."

Hughes' description, which is typical of the analyses presented in the examples of this approach considered in this paper [8, 9, 22, 23], makes the following technical points:

- Every node in the tree represents either a speciation or a duplication. It is possible to find the set of duplication nodes by comparing the gene family tree to a species tree such as the cartoon of the Tree of Life shown in Figure 2. Hughes identified two duplication nodes (14 and 15). There are two more duplication nodes in the RXRB clade (3 and 6) that he does not mention.

- Bounds on the time of duplication can be inferred for each duplication node from the relative positions of speciation and duplication nodes in the tree. According to the topology shown in Figure 1, duplications 14 and 15 are both bounded above by the divergence of vertebrates and insects and bounded below by the divergence of tetrapods and bony fishes. The upper bound can be inferred from the clustering of insect genes outside the gene family clades and the lower bound from the presence of a fish gene in each subfamily clade.

- When a duplication hypothesis depends on a node with weak support in the sequence data, alternative hypotheses should be considered. Because the bootstrap values associated with the zebrafish branches in Figure 1 are low, topologies in which zebrafish genes do not cluster within the subfamilies should also be considered. For this reason, the divergence of the amphibian lineage may be a more reliable lower bound for duplications 14 and 15.

Our Results. Hughes' analysis is typical of many studies in the biology literature of gene duplication using ad hoc analysis of gene family trees [2, 8, 9, 11, 22, 23]. Unlike the RXR example, many of these trees are unrooted because it is frequently not possible to find a sequence from the gene family in a suitable outgroup species. A rooted GFT is a hypothesis concerning the evolutionary history of a gene family from which the number of duplications that occurred, a partial ordering on their occurrence and a time range for each duplication can be inferred. An unrooted gene family tree represents a set

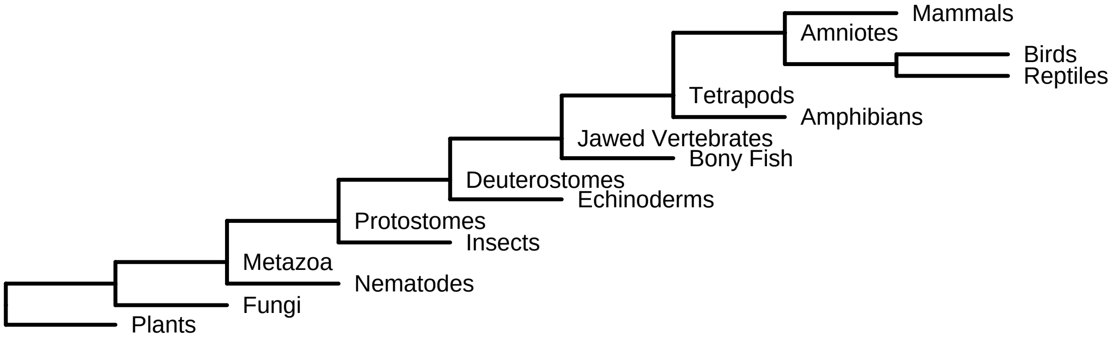

Figure 2: A species tree showing major speciation events in the eukaryote lineage. This tree was derived from the University of Arizona Tree of Life project [13] and the NCBI Taxonomy database [16]

of such hypotheses, one for each possible rooting. Three problems of reconstructing the duplication history of a gene family can be stated formally as follows:

> **Rooted Trees:** Given a rooted GFT, G, and a species tree for the species represented in G, identify all duplication nodes and determine lower and upper bounds on the time of each duplication.
>
> **Unrooted Trees:** For every possible rooting of an unrooted GFT, G, determine the duplication history of that rooting.
>
> **Alternate hypotheses:** Given a threshold for minimum, acceptable bootstrap support and a rooted GFT, G, determine the set, W, of all edges with bootstrap support below the threshold. Generate the set of alternate rooted GFT's that can be obtained using local rearrangements around edges in W. Reevaluate the duplication history accordingly.

After reviewing previous work involving the relationship between gene family trees and species trees (Section 2), we present algorithms for automatically determining the history of duplications in rooted GFT's in Section 3. In Section 4, we present an algorithm for computing all duplication histories of an unrooted tree in linear time and discuss criteria for evaluating these alternate hypotheses. The generation and evaluation of alternate duplication histories for weak edges is discussed in Section 5.

We implemented these algorithms and tested them on gene family trees published in the molecular evolution literature [8, 22, 23]. As summarized in Section 6, the hypotheses for rooted trees generated by our program are consistent with the assessments presented in the original papers. For unrooted trees and nodes with low bootstrap values, our program generates and scores all alternate hypotheses, providing an exploratory analysis tool. In addition, an explicit statement of all hypotheses helps mitigate any biased expectations of the data the user might have.

2 Related Work

The problem of disagreement between gene trees and species trees was first raised by Goodman *et al.* [6] in the context of inferring a species tree from a gene tree that may contain paralogies. They introduced the notion of a map between a gene tree and a species tree and suggested a cost function for evaluating a species tree with respect to a gene tree based on edit distance, gene duplication and gene loss.

These concepts were further developed and formalized in [7, 15, 19, 20, 26]. Formally, given a rooted gene tree, G, the problem is to find the species tree, T, that optimizes an evaluation criterion. Several optimality criteria have been proposed (see [3, 4] for a comparative survey), all of which attempt to capture the notion that gene duplication and subsequent loss are rare events. These criteria involve constructing a mapping, $M : G \mapsto T$, which is used to compute the cost function. Several authors have pointed out that it is difficult to distinguish true gene loss from genes that have not yet been sequenced and discuss approaches to distinguishing true losses from apparent losses in the cost function [6, 15, 20].

When inferring a species tree from a gene tree, the gene tree is assumed to be correct and the true species tree is unknown. The problem of finding an optimal species tree is NP-hard [29] for the optimality criteria considered so far. In contrast, we assume that the true species tree is known and use it to infer the duplication history of a gene tree. While we share some mathematical structure with [7, 15, 19, 26], most notably the mapping M, we consider the problem of dating duplication events and generating and evaluating alternate hypotheses. Dating duplication events in rooted and unrooted trees is a computationally tractible problem, which is crucial if we hope to apply this to large data sets.

The methods to infer species trees from gene trees surveyed here do implicitly generate duplication histories in rooted trees although the time of duplication is generally not considered. In addition, most optimality criteria surveyed here are subject to the constraint that each species may only be represented once in G and hence would not be suitable for our application. A notable exception is the work of Page and Charleston [21], who have developed two software packages, Component and GeneTree, that, as well as inferring species trees, will compute and display duplication histories for rooted gene trees. This provides an interactive, exploratory analysis tool, but could not be used to automate the analysis of large data sets. None of the work surveyed addresses alternate rootings of unrooted trees or alternate hypotheses due to weak edges. These two problems are addressed in the current paper.

3 Rooted Trees

In general, only a subset of the descendants of a duplication event will appear in a GFT, either because some paralogous sequences have been lost due to mutation or because they have not yet been sequenced. The problem of determining a duplication history in the absence of a complete GFT is exemplified in Figure 3, which shows two alternate phylogenies for a hypothetical gene family, A, with subfamilies, A_1 and A_2. Figure 3(a) suggests that gene A arose before the divergence of fish and tetrapods and was duplicated after the divergence of fish and before the separation between birds and mammals. In contrast, Figure 3(b) implies that the duplication took place before the divergence of fish and tetrapods. Although there is only one fish sequence, it clusters with the genes in the A_2 family, suggesting either that the fish A_1 gene has been lost due mutation or deletion or that it has not yet been sequenced. We exploit this intuition to obtain an algorithm to identify and date duplication nodes in a rooted GFT.

Let S be a set of orthologous and paralogous gene sequences from a gene family; G, a binary phylogeny inferred from the sequences in S; and T, a binary species tree containing the species in S. Both the identification of duplication nodes and the calculation of duplication dates requires constructing a mapping, M, from every node in G to a target node in T. Let n be a node in G and let $l(n)$ and $r(n)$ be its left and right children, respectively. M maps each leaf node in G to the node in T representing the species from which the sequence was obtained. (Leaf nodes in G represent sequences, whereas leaf nodes in T represent species.) Each internal node in G is mapped to the least common ancestor (lca) in T of the target nodes of its children; that is, $M(n) = lca(M(l(n)), M(r(n)))$. For example, in Figure 3(b), the leaf nodes are mapped to *chicken*, *human*, *fish*, *chicken*, *mouse*, from top to bottom. $M(x)$ = *amniote*, since the lca of *mouse* and *chicken* is *amniote* in the Tree of Life (Figure 2). $M(z)$ is also *amniote*, while y and w both map to *jawed_vertebrate.*

An algorithm for constructing the mapping, M, and identifying duplication nodes has been developed in the context of using multiple gene trees to generate a species tree. By using fast lca queries [10][1], M can be computed in linear time. While our goals are different, we share a key algorithmic component with this work. We refer the reader to [15] for a complete description and proofs.

Observe that under the mapping, a node n in G is a speciation node if its children are mapped to independent lineages in T. In Figure 3(b), x is a speciation node since mammals and birds are separate lineages. If the children of $M(n)$ share a lineage, then n is a duplication node. When this occurs, one child's target in T is an ancestor of the other's and n will be mapped to the same label as the ancestral child. For example, node w is a duplication node in Figure 3(b) because $M(y)$ = *jawed_vertebrate* is an ancestor of $M(x)$ = *amniote.*

Observation 1 *Node n is a duplication node if and only if $M(n) = M(l(n))$ or $M(n) = M(r(n))$.*

The mapping, M, can also be used to compute lower and upper bounds on the time of duplication. Let n be a duplication node in G. Since copies of the duplicated gene are observed in descendents of both $l(n)$ and $r(n)$, the duplication must have been present in their last common ancestor, yielding the lower bound $L(n) = M(n)$. By a similar argument, the upper bound can be shown to be the target of the nearest ancestor, a_n, of n that is a speciation node. Since copies of the duplicated gene are present in only one of the subtrees rooted at children of a_n, the duplication must have occurred in a more recent species. If n has an ancestor that is a speciation node, we set $U(n) = M(a_n)$. Otherwise, $U(n)$ is the origin of life. For example, in Figure 3(b), the bounds on the duplication node, w, are $L(w)$ = *jawed_vertebrate* and $U(w) = \infty$, since w is the root node of G. In Figure 3(a), b is a duplication node with label *amniote.* Its parent, a is a speciation node with label *jawed_vertebrate.* Thus, $L(b)$ = *amniote* and $U(b)$ = *jawed_vertebrate* .

[1] Several early papers on lca computation were too complicated to implement, even papers which claimed to be "simplifications", and had large hidden constants. Thus, it is a "folk theorem" that any algorithm which uses lca precomputation is impractical. However, the state of the art of lca computation has progressed since those early papers, and there now exist lca algorithms which are very simple and very practical.

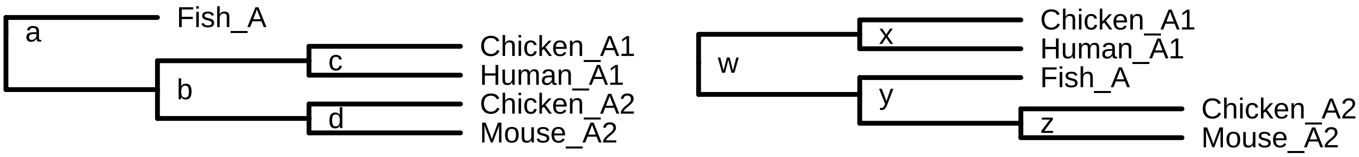

Figure 3: Two examples of rooted gene family trees representing two alternate hypotheses of the evolution of the fictional gene A family.

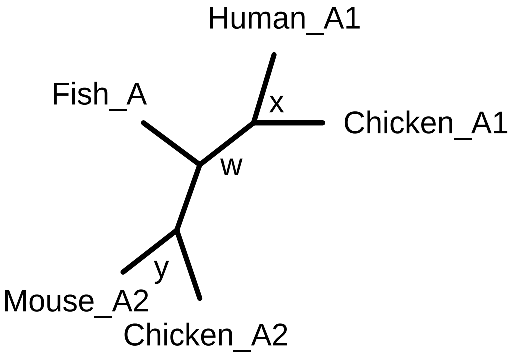

Figure 4: An example of an unrooted gene family tree for the gene A family.

Observation 2 *The duplication associated with a node, n, in G, occurred after the speciation event $U(n)$ and before the speciation event $L(n) = M(n)$.*

Notice that we can compute U in linear time.

4 Unrooted Trees

The rooted tree in Figure 3 allows us to observe evidence of the duplication through clustering, even when one of two paralogs is missing. In contrast, the unrooted tree in Figure 4 shows that without a root, it is impossible to tell whether the duplication in the A family took place before or after the evolution of fish. If the tree in Figure 4, is rooted on the edge ($Fish\_A$, w), then we hypothesize that the duplication occurred after the evolution of fish. If we root the tree on edge (w,y) (or (w,x)), then the duplication occurred before the evolution of fish and $Fish\_A$ is a member of the A_1 (or A_2) subfamily. The hypotheses associated with rooting the remaining edges seem unlikely since they require three duplications and substantial gene loss.

Unlike a rooted tree, which encodes a single evolutionary hypothesis, an unrooted tree with $|E|$ edges represents up to $|E|$ different hypotheses, one for each possible rooting. Given an unrooted GFT G, we wish to label each node in G as either a duplication or speciation node under every possible rooting. A simple quadratic time algorithm would be to apply the rooted tree algorithm to every possible rooting. However, we can derive a linear time algorithm as follows. Notice that, with respect to a node v, we can partition all possible rootings of the tree into 3 groups: the root must be in one of three directions, according to which of the edges incident on v is on the path from n to the root. Let e_1, e_2 and e_3 be the edges incident on v. The status of v as either a duplication or speciation only depends on which edge points towards the root. This is because if we fix which edge is up, the subtree rooted at v is fixed, and so is the bottom-up lca computation. The point is that we need now only compute $M_{e_1}(v)$, $M_{e_2}(v)$ and $M_{e_3}(v)$ – one $M(\cdot)$ value for each possible "up" edge, from which we can compute the labeling under any desired rooting in linear time.

To compute the three values we simply do the recursive computation at each node in any order. That is, suppose we want to compute $M_{e_2}(v)$, for some v. This determines which two nodes are down. Call them u and w. Then $M_{e_2}(v) = lca(M_{\{v,u\}}(u), M_{\{v,w\}}(w))$. We recursive compute $M_{\{v,u\}}(u)$ and $M_{\{v,w\}}(w)$. In order to keep from recomputing the same value over and over, we simply store all values in a table as we compute them. Thus, once we have computed $M_{\{v,u\}}(u)$ once recursively, we can look it up in constant time without need for recomputation in the future. Thus, all $3n$ values can be computed in $O(n)$. Note that we can incorporate the GNNI heuristic discussed in Section 5 during the recursive computation without increasing the complexity of the algorithm.

Having generated a duplication history for each possible rooting, we wish to score the alternate hypotheses so that the results can be presented in the order most worthy of further scrutiny. A scoring function implicitly represents an evolutionary model concerning the processes of speciation, duplication and gene loss. The user should be able to select the scoring function (and hence, the model) best suited to the data set and the question to be investigated. The use of a scoring function to rank duplication histories is similar to inferring a species tree from a gene tree with respect to an optimality criterion, since the plausibility of a duplication history is evaluated in both cases. However, those optimality criteria that can only applied to gene trees in which each species is only represented once cannot be used here. In the current work, we present one scoring

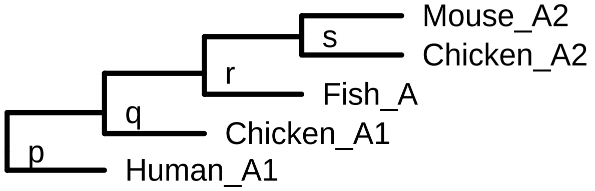

Figure 5: An unlikely rooting for the GFT for the gene A family.

function as a proof of concept.

Let M^* be the label in T of the lca of the set of species in S; that is, M^* is the root of T. Define the cost, $C(G)$, of a rooted GFT G, to be the number of *duplication* nodes, n, in G such that $M(n) = M^*$. The cost of a tree without paralogies is always zero.

The original motivation for this scoring function was the observation that high labels in G (that is, labels close to M^*), tend to "trickle up" the tree. This is because, given nodes x and y in G, if y is an ancestor of x, then $M(y) \geq M(x)$. In particular, if $M(x) = M^*$ then all nodes ancestral to it must also be labeled M^* and all must be duplication nodes. High labels also tend to propagate up the tree and force duplication nodes.

As an example, consider the unrooted tree in Figure 4. For this tree, $M^* =$ *jawed_vertebrate*. The trees in Figure 3 are two plausible rootings for this tree, with costs of zero and one, respectively. Each implies a single duplication. In contrast, consider the rooting shown in Figure 5. This rooting has a higher cost. The internal node r is the lowest node in the tree to be labeled with $M^* =$ *jawed_vertebrate*, forcing $M(p) = M(q) = M^*$ as well. This rooting also yields a duplication history with two duplications and substantial gene loss, since it implies that one copy from the first duplication was lost from (or is as yet unsequenced in) all taxa except human and one copy from the second duplication was lost in all taxa except chicken. Thus a cost function based on a mathematical observation, the "trickle up effect", implies an evolutionary model: duplication and gene loss are rare events. Note that this cost function can be used to compare alternate rootings of the same tree but costs of two different trees cannot be compared, since the minimum cost depends on the structure of the tree.

5 Rooted Tree Rearrangements

As first suggested by Goodman *et al.* [6], the history of a gene family should ideally be inferred using an evolutionary model that takes sequence evolution, gene duplication and gene loss into account, but it is not obvious how to combine these different types of information. Our approach is to start with a tree inferred from sequence alone, but to use a model of gene duplication and loss to consider alternate hypotheses for edges that are not strongly supported by the sequence data.

A measure of confidence can be associated with every edge in a phylogeny using bootstrapping [1]. Every edge, e, in a tree bipartitions the set of leaf nodes. If the bootstrap value of e is low, it suggests that the evidence in the data for that bipartition is weak. It does not reflect on the certainty of the structure of any other part of the tree. In reconstructing the duplication history of a rooted GFT, we consider alternate hypotheses associated with a weak edge, e, by generating *Nearest Neighbor Interchanges (NNI's)* around e. This rearrangement [27] generates alternate bipartitions for e while leaving all other bipartitions associated with the tree unchanged.

An NNI will change the mapping, $M(\cdot)$, resulting in a new mapping, $M'(\cdot)$. In some cases, this will also change the duplication history. Figure 6(a) shows a tree fragment with two internal nodes both labeled *vertebrate*. NNI a' leaves the labeling unchanged. However, NNI a'' changes the label of the deeper internal node, thereby eliminating a duplication. In Figure 6(b), one rearrangement again leaves the mapping unchanged. The other rearrangement (b'') changes $M(\cdot)$ and moves the duplication to the deeper node.

As in the case of unrooted trees, the tree that represents the best hypothesis for the duplication history of the gene family can be selected with respect to an optimization criterion, such as the function $C(\cdot)$ defined in the previous section. Formally, let W be the set of edges with bootstrap values below a threshold provided by the user and let $\mathcal{G}_W$ be the set of trees that can be derived from G by NNI operations across edges in W. Given a GFT G, a set of weak edges W, and a species tree T, find the tree $G' \in \mathcal{G}_W$ such that G' optimizes some criterion.

As a heuristic, rearrangements associated with individual edges can be evaluated, and accepted or rejected, independently for each edge. Below we describe a heuristic for deciding whether to accept a rearrangement around a weak edge.

> **Greedy NNI (GNNI):** Let $e = (x, y)$ be a weak edge in a rooted tree, G, where x is a descendent of y. Let $M(x)$ be the label of x and $M'(x)$ the label of x after an NNI rearrangement. (We call x the *pivot* of the NNI rearrangement.) Perform an NNI if $M'(x)$ is a strict descendent of $M(x)$ in T.

This heuristic is based on the "trickle up" effect: when nodes in G are incorrectly mapped to labels high in T, false duplication nodes can result. GNNI attempts to eliminate such false duplications by accepting rearrangements that remap the pivot to a lower node. It will accept such arrangements even if it causes the pivot to be converted from a speciation node to a duplication

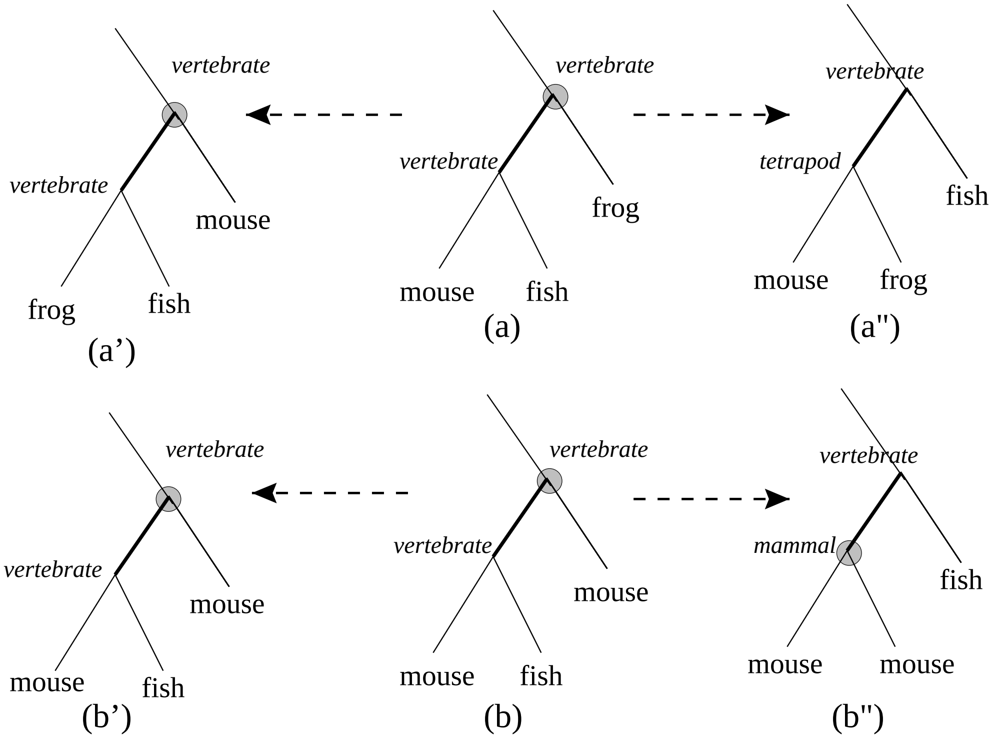

Figure 6: Two tree fragments, each with the three possible Nearest Neighbor Interchanges around the edge shown in bold. Duplication nodes are shown as grey circles.

node, the logic being that this change may eliminate false duplications further up the tree.

In Figure 6(a), we present a scenario where the frog and fish genes were incorrectly placed with respect to each other due to weak signal in the sequence data. Since NNI (a'') lowers the label at the pivot, the greedy heuristic corrects this error, eliminating a false duplication node. The example in Figure 6(b) is more ambiguous. The middle tree represents a duplication before the divergence of fish followed by a loss in the fish lineage. Figure $6(b'')$ represents a duplication after the fish-tetrapod split. Which scenario is more likely requires specialized knowledge of the processes of duplication and loss and probably depends on the specific properties of gene family as well. GNNI will select $6(b'')$ because it reduces the label of the pivot from *vertebrate* to *mammal.*

Like the cost function, $C(\cdot)$, GNNI implies a hidden evolutionary model: by moving duplications towards the leaves of the tree, it has the effect of selecting hypotheses with fewer duplications and losses. It also encourages more recent duplications. Notice that GNNI does not take global properties of G and G' into account. If several edges are rearranged in succession, the order in which they are visited may affect the tree ultimately obtained. Since GNNI is based on the "trickle up" effect, it should be applied bottom up.

6 Experimental Results

The algorithms described in the previous section have been implemented in Java program called NOTUNG. NOTUNG takes a gene family tree, G, a species tree, T and a bootstrap threshold, τ, as input. Input trees are represented in Newick format [17]. For rooted trees, NOTUNG generates a gene duplication history as output; that is, a list of duplication nodes, with bounds on the time of duplication for each one. NOTUNG also applies the Greedy NNI heuristic to edges with bootstrap value less than τ in bottom up order, generating an alternate tree, G', if rearrangements at any of the edges are accepted. In this case, it also presents the duplication history for G' and the list of node swaps that converted G to G'. For unrooted trees, NOTUNG considers all possible rootings and computes a duplication history for each. These histories are ranked according to the cost function, $C(\cdot)$, presented in Section 4. Note that NOTUNG, as designed, works equally with binary and higher degree trees, though the exact implementation of NNI heuristics in a tree with degree greater than two is somewhat problematic, both in terms of increased computation time, and in terms of generating biologically reasonable heuristics.

Our intent is to provide an exploratory analysis tool that allows the user to review all alternate hypotheses. Heuristics are used to suggest which alternatives are most worthy of attention. One goal of the experimental work presented below is to determine whether our heuristics rank alternative hypotheses effectively. Below we describe NOTUNG'S performance on rooted trees, unrooted trees and trees with low bootstrap values. As test data, we used all "non-pathological" trees from three recent articles on large scale duplication [8, 22, 23]. We eliminated non-binary trees and trees based on genes with complicated internal structure such as mosaic genes or genes with repeated domains, and trees that show evidence of horizontal gene transfer. We analyzed the remaining thirteen trees using NOTUNG and compared the automatically generated results with the verbal analysis presented in the source paper.

The program compares the input GFT with a species tree to infer the duplication history. Since there are many competing hypotheses concerning the topology of the Tree of Life, our program allows the user to supply a species tree as input. In the experiments described below, we tried, to the extent that it was possible to determine from the text, to use the same Tree of Life as the authors who originally analyzed the tree. Most authors used a tree consistent with that shown in Figure 2. Pebusque *et al.* [22] used a variant in which nematodes are included in the Protostome clade. Our standard species tree is constructed from information in the University of Arizona Tree of Life project [13] and the NCBI Taxonomy database [16].

6.1 Rooted Trees

The rooted trees in our data set, representing the NOTCH[2], RXR, C, PBX, TEN and HSP70 gene families, were originally presented by Hughes [8]. The histories constructed by the program were consistent with Hughes' analysis in all cases. Generally, NOTUNG finds a superset of the duplications discussed by Hughes, since he only mentions those duplications that are relevant to the biological question he is addressing. This was true of all the trees reported here; the authors of the original studies did not attempt to describe the entire duplication history. They simply reported the aspects they considered relevant to their research. In contrast, NOTUNG reports the entire history, including variants, and allows the user to triage the information.

As an example, we show the duplication history generated by NOTUNG for the RXR tree shown in Figure 1:

```
Cost = 0
Duplication at 15  LB: jaw        UB: pro
Duplication at 14  LB: jaw        UB: pro
Duplication at 6   LB: dani_reri  UB: jaw
Duplication at 3   LB: xeno.laevi UB: tet
```

Here "jaw" refers to *jawed_vertebrate*, "pro" to *protostomes* and "tet" to *tetrapods*. Both duplications occurred after the divergence of protostomes (insects and molluscs) from deuterostomes (fish and tetrapods), which is consistent with Hughes' analysis. It also find the more recent duplications not discussed by Hughes.

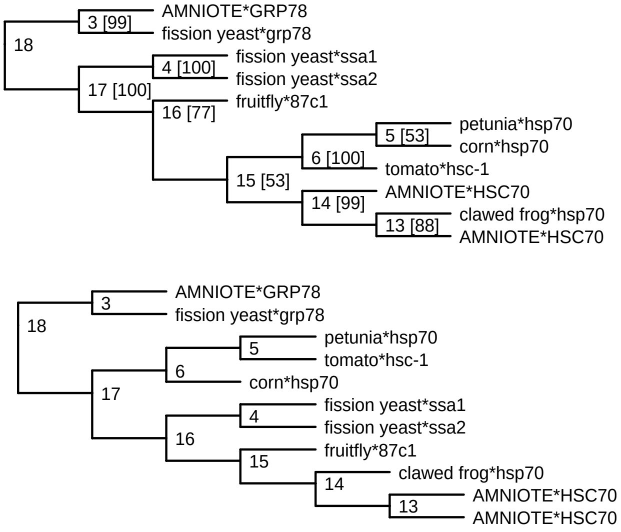

Figure 7: The HSP [8] tree before and after NNI rearrangements. The trees have been simplified by compressing clades containing only mammals and birds (AMNIOTE*GRP78m, AMNIOTE*HSP70, AMNIOTE*HSC70). No rearrangements were accepted in these clades. Internal nodes are given numerical labels. In the upper tree, the bootstrap values of edges with bootstrap support above 50 are labeled with square brackets.

6.2 Alternate Hypotheses for Weak Branches

Alternate hypotheses were evaluated for every branch with a bootstrap value less than 90% in the six trees described in the previous section. Rearrangement trees were generated for three of them. In the remaining trees, no NNI's were accepted under the greedy heuristic. All accepted rearrangements fell into the two categories described in Section 5: phylogenetic corrections (e.g., Figure 6(a)) and more controversial alternate hypotheses characterized by more recent duplications and fewer gene losses (e.g., Figure 6(b)).

Both types of rearrangement appear in the HSP70 trees shown in Figure 7. These trees have been simplified for the purposes of exposition. Subtrees containing only birds and mammals have been compressed and are shown in capital letters. The upper tree shows the original topology before rearrangements were considered. This tree contains five branches with bootstrap values below the threshold. Two of them are adjacent. Initially, our program inferred a duplication history with

[2] A rat sequence was removed from the NOTCH tree to obtain a binary tree.

nine duplications (nodes 4, 6, 8, 10, 12, 14, 16, 17 and 18) and a score of three (nodes 16, 17 and 18 were labeled $M^* = eukaryotes$). The structure of this tree, (fungi (insects (plants, vertebrates))), is at odds with the structure of the Tree of Life, (plants (fungi, animals)). The structure within the plant clade also disagrees with the Tree of Life since petunias and tomatoes are more closely related to each other than either is to corn.

In contrast, the topology after rearrangement (the lower tree) had three fewer duplication nodes and a score of one. Duplications at nodes 6, 16 and 17 were eliminated and 14 was replaced by 13. The removal of duplications from nodes 6, 16 and 17 can be interpreted as correcting errors in the original topology. That topology implies that an ancestral HSP gene was duplicated twice early in the eukaryote lineage; subsequently each of the four resulting copies survived in only one lineage (fungi, insects, plants and vertebrates, respectively) and was lost in the other three. In view of the low bootstrap support, it seems more plausible that the yeast and fly sequences are placed incorrectly. In the rearranged tree, the branching of plants, yeast and insects is compatible with the Tree of Life. This second hypothesis is more compelling than the original hypothesis of two early duplications followed by massive gene loss. The exchange of the corn and tomato genes to remove the duplication at node 6 also appears to correct an error in the reconstruction of the tree topology. The rearrangement of the frog sequence that led to the replacement of the duplication at node 14 with one at node 13 is more controversial. It is open to interpretation whether a duplication in the amniote lineage is more or less likely than a duplication before the divergence of amphibians followed by loss of one copy.

Several aspects of the NNI method are illustrated by this example. First, rearrangement can result in substantially different hypotheses. The number of duplication nodes in the rearranged HSP tree decreased from nine to six. As this illustrates, although it is possible to pick out individual rearrangements of low confidence branches by eye, when a tree contains many weak branches it is helpful to have a tool to integrate all the alternate hypotheses automatically. Second, when weak branches are adjacent, the order in which NNI's are applied matters. If NNI were not applied bottom up, the rearrangements leading to the elimination of duplication nodes 16 and 17 would not have been accepted.

6.3 Unrooted Trees

We tested Notung on seven unrooted trees: the TCF, CRYB and LIM families [23] , the VMAT, ANK and EGR families [22] and the PSMB family [8]. For each tree, Notung computed the duplication history for every root and ranked them according to the cost function, $C(\cdot)$. We compared this ranking with the rootings favored by the authors.

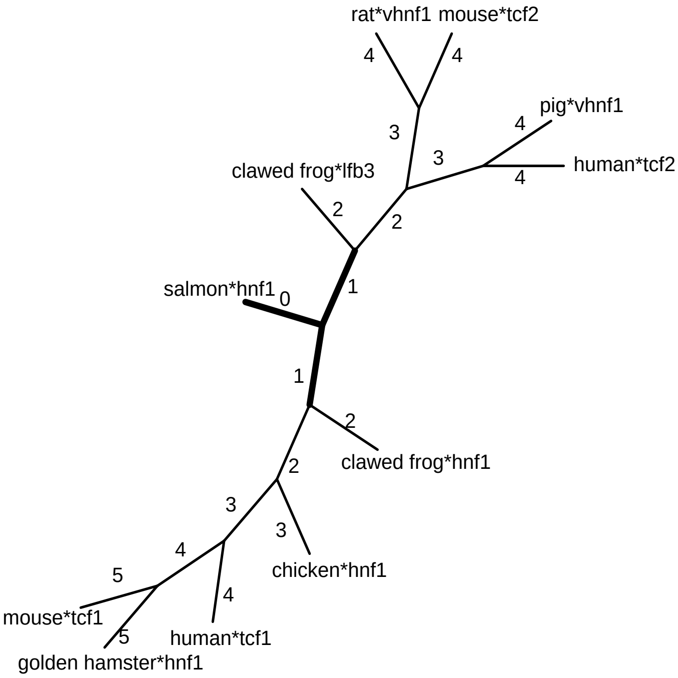

Figure 8: An unrooted tree for the TCF family [23]. Each edge, e, is labeled with the cost, $C(\cdot)$ of the tree rooted at e. Edges in bold correspond to rootings supported by the analysis in [23].

Although possible rootings are rarely, if ever, mentioned explicitly by the author's whose trees we tested, they frequently imply that only a subset of the possible rootings lead to plausible hypotheses. Consider, for example, the TCF family tree, shown in Figure 8 with rooting scores labeling each edge. In their analysis, Ruvinsky and Silver [23] state that “it is difficult to conclude whether the split between the TCF1 and TCF2 subfamilies occurred before or after the separation between fish and tetrapods,” but “in any case, divergence between the two sub families has taken place prior to the amniote-amphibian separation.” These conclusions are consistent with a rooting on the bold edges in Figure 8 and no others.

For each tree, we partitioned the set of edges into plausible and implausible rootings from the analysis presented by the original authors and compared this partition with the output of Notung. For five out of the seven trees, all plausible rootings ranked above all implausible rootings. For the remaining two trees, the costs of all implausible rootings were greater than or equal to the costs of all plausible rootings. For one of these, the PSMB tree, the set of highest ranked edges is a superset of the rootings deemed plausible by Hughes.

One of these edges has weak bootstrap support. When the NNI heuristic was applied to this edge, a rearrangement was accepted according to the greedy criterion. When the rearranged tree was rescored, the set of lowest cost edges exactly agreed with Hughes' analysis. In the other case, the CRYB tree, there were eight top-ranked rootings of equal cost, while the authors' analysis implied that only one rooting is possible. The duplication histories (i.e., the set of duplication nodes with time ranges) were identical for the eight edges. Only the ordering of the duplication nodes differed. This suggests either that the authors did not consider all alternate scenarios, possibly missing something of interest, or that they had additional information about the gene family, such as the biochemical properties or functional roles of the proteins, that allowed them to rule out other rootings.

Within the set of plausible rootings, the ordering of scores does not always agree with the biologists' assessments. In contrast to the analysis of Ruvinsky and Silver, who ranked the three best edges equally, the edge adjacent to the fish sequence is ranked higher than the other two because the scoring function favors more recent duplications. In fact, this decision should reflect an evolutionary model explicitly chosen by the user.

6.4 Discussion

In this study, we analyzed every non-pathological tree in three papers [8, 22, 23]. The duplication histories generated and the rankings of alternate rootings were consistent with the analyses of the authors of the original papers for all trees considered. This confirms that NOTUNG is a useful exploratory data analysis tool. The cost function used to rank alternate rootings correctly identified unlikely hypotheses, providing the user with a way to control the quantity of output to be reviewed. For edges with low bootstrap values, the GNNI heuristic was effective in correcting errors in the duplication history stemming from errors in the original tree topology. It also identified more controversial alternatives. While these are of interest and should be presented to the user for consideration, it would be useful to be able to separate likely and speculative rearrangements. Since these are very simple heuristics, we are confident that with further experimentation and better models of gene duplication and loss, improved evaluation methods for duplication histories and rearrangements can be developed.

Currently, there is a great deal of interest in using gene duplications to study the role of whole genome duplications in genome evolution [25, 28]. This will require dating all paralogs in a genome. In its current form, NOTUNG can be used to estimate duplication dates of rooted GFT's automatically. With more reliable evaluation methods, NOTUNG can also be adapted to the automatic analysis of unrooted trees and rearrangements of trees with weak edges.

Acknowledgements

The authors wish to thank Jim Brown, Ilya Ruvinsky, Lee Silver and Mona Singh for helpful discussions. Species trees in Newick format were downloaded from the Ribosomal Database Project [14]. Most figures in this paper were drawn with Treetool, an interactive tree-plotting program written by Mike Maciukenas, for the Ribosomal Database Project [14].

References

[1] Bradley Efron and Gail Gong. A leisurely look at the bootstrap, jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[2] T. Endo, T. Imanishi, T. Gojobori, and H. Inoko. Evolutionary significance of intra-genome duplications on human chromosomes. *Gene*, 205(1-2):19–27, 1997.

[3] O. Eulenstein, B. Mirkin, and M. Vingron. Comparison of a annotating duplication, tree mapping, and copying as methods to compare gene trees with species trees. *Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:71–93, 1996.

[4] 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.

[5] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):563–7, 1996.

[6] M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera, and G Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*, 1979.

[7] R. Guigo, I. Muchnik, and T.F. Smith. Reconstruction of ancient phylogenies. *Molecular Phylogenetics and Evolution*, 6:189–213, 1996.

[8] A. L. Hughes. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *MBE*, 15(7):854–70, 1998.

[9] A. L. Hughes. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *JME*, 48(5):565–76, 1999.

[10] Joseph JáJá. *Introduction to Parallel Algorithms.* Addison-Wesley, Reading, MA, 1991.

[11] M. Kasahara. New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas*, 127(1-2):59–65, 1997.

[12] R. C. King and W. D. Stansfield. *A Dictionary of Genetics.* Oxford University Press, 1990.

[13] D. R. Maddison and W. P. Maddison. Tree of life. *http://phylogeny.arizona.edu/tree/phylogeny.html.*

[14] B. L. Maidak, J. R. Cole, C. T. Parker, G. M. Garrity, N. Larsen, B. Li, T. G. Lilburn, M. J. McCaughey, G. J. Olsen, R Overbeek, S Pramanik, T. M. Schmidt, J. M. Tiedje, and C. R. Woese. A new version of the rdp (Ribosomal Database Project). *Nucleic Acids Res*, 29(1):171–3, 1999.

[15] B. Mirkin, I. Muchnik, and T.F. Smith. A biologically consistent model for comparing molecular phylogenies. *Journal of Computational Biology*, 2:493–507, 1995.

[16] The NCBI Taxonomy database. *http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html.*

[17] The Newick tree format. *http://evolution.genetics.washington.edu/phylip/newicktree.html.*

[18] S. Ohno. *Evolution by Gene Duplication.* Springer-Verlag, 1970.

[19] R.D.M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. *Syst Zool*, 1994.

[20] R.D.M. Page and M.A. Charleston. Reconciled trees and incongruent gene and species trees. *Mathematical Heirarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:57–70, 1996.

[21] R.D.M. Page and M.A. Charleston. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240, 1997.

[22] M.-J. Pebusque, F. Coulier, D. Birnbaum, and P. Pontarotti. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *MBE*, 15(9):1145–59, 1998.

[23] I. Ruvinsky and L. M. Silver. Newly indentified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a t-box cluster duplication. *Genomics*, 40:262–266, 1997.

[24] L. M. Silver. *Mouse Genetics.* Oxford University Press, 1995.

[25] L. Skrabanek and K.H. Wolfe. Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev*, 8(6):559–565, 1998.

[26] U. Stege. Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. In *Proceedings of the 6th International Workshop on Algorithms and Data Structures (WADS'99)*, 1999.

[27] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogeny inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates Inc., Sunderland, MA., 1996.

[28] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.

[29] L. Zhang. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4:177–188, 1997.