Notung: Dating Gene Duplications using Gene Family Trees

Kevin Chen,[*]   Dannie Durand[†]   Martin Farach-Colton[‡]

September 30, 1999

Abstract

Gene duplications are a widely studied phenomenon. Gene duplications differ from other genomic rearrangments, such as transpositions and reversals, in that the time of duplication can be estimated; that is, we can in some cases calibrate duplications with respect to speciation events. The dating of duplication events has been used to argue for or against hypotheses of large scale genomic duplication. For example, if we know that one gene duplication occurred before some speciation event, and some gene duplication occurred after that speciation event, then we can assume that the two genes were not copied in a single large scale duplication.

Many studies of gene duplication dating have appeared in the molecular evolution literature. These analyses have been performed by hand. Data sets are growing substantially, so that they are now at the limit of what can be analyzed by traditional methods, and even existing data sets admit alternative hypotheses which would be too tedious to consider without automation. In this paper, we provide a toolbox called NOTUNG with which to analyze gene duplication events. NOTUNG yields results which are consistent with all duplication event dating papers surveyed, as well as generating plausible alternative hypotheses in many cases. Thus NOTUNG provides a basic building block for analyzing the history of gene duplications.

[*]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*). Supported by an Alfred P. Sloan Computational Molecular Biology Fellowship.

[‡]Department of Computer Science, Rutgers University, Piscataway, NJ 08855, USA (*farach@cs.rutgers.edu, http://www.cs.rutgers.edu/∼farach*). Supported by NSF Career Development Award CCR-95-01942, NSF Grant BIR-94-12594, an Alfred P. Sloan Research Fellowship and NATO Grant 96-0215.

1 Introduction

Yeast is a single cell organism with 6000 genes [6], while mice have an estimated 50,000 - 100,000 genes [27]. How did this order-of-magnitude increase in gene number, with its concomitant increase in functional complexity, arise? Gene duplication followed by differentiation of sequence and function through mutation is posited to be a primary mechanism for acquisition of new genetic function in the evolution of increasing organismal complexity [21], though the exact role of duplications in genomic evolution is widely debated in the biology literature.

For example, by combining phylogenetic and developmental data, Ruvinsky and Silver [24] have presented strong evidence linking the duplication of a pair of regulatory genes involved in limb development to the doubling of the number of limbs found in early vertebrates, suggesting a mechanism for the evolution of tetrapods. On a larger scale, Ohno hypothesized that two whole genome duplications occurred early in vertebrate evolution [21]. Ohno's theory is quite controversial, and many alternate hypotheses for large scale duplication in vertebrates have been presented (see [28] for a survey). Whole genome duplications are also thought to have played an important role in the evolution of the maize [19] and yeast [29] genomes.

The increasing availability of whole genome sequence data is raising new computational problems in elucidating the historical, spatial and functional relationships between genes in genomes. In this paper, we present a computational method for dating gene duplications for use in analyzing one such problem: reconstructing the role of gene duplication in genome evolution. Our results are a set of tools called NOTUNG which can be used for exploring alternative hypotheses about duplication events. We therefore begin by considering a typical "by hand" analysis of a set of duplications.

An Example of Duplication Analysis. In particular, we review Hughes' analysis of the evolution of the RXR family [10]. His analysis is based on the rooted tree reproduced in Figure 1. Hughes constructed the tree from gene sequences from five vertebrate and three invertebrate species using the Neighbor Joining heuristic. Confidence in clustering patterns was assessed using bootstrapping, a statistical resampling method. 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 as strong (99%) boot strap 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' verbal description, which is typical of the analysis presented in of the examples of this approach [10, 11, 23, 24] cited here, 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.

There are many examples of this approach in the biology literature [4, 10, 11, 13, 23, 24]. In these articles, approximate time ranges for gene duplications are determined using ad hoc analysis of gene family trees. A gene family is "a set of genes descended by duplication and variation from some ancestral gene" [14], 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). Such a tree represents a hypothesis concerning the evolution of the gene family, showing the relationship between speciation and gene duplication events. It differs from a species tree in that a species may appear more than once in the tree. Gene family trees may be rooted or unrooted, depending on whether a sequence from the gene family is available in a suitable outgroup species.

Our Results. In this paper, we automate the process of inferring a duplication history from a gene family tree. A rooted gene family tree 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 of such hypotheses, one for each edge. The problem of reconstructing the duplication history in a rooted gene family tree can be stated formally as follows:

> Given a gene family tree, a species tree for the species represented in the gene family tree and a threshold for minimum, acceptable bootstrap support,
>
> - **Duplications:** identify all duplication nodes,
>
> - **Time bounds:** determine lower bounds, and upper bounds when possible, on the time of each duplication,
>
> - **Alternate hypotheses:** for each branch with bootstrap support below the threshold, generate an alternate gene family tree by considering local rearrangements of that branch. Reëvaluate the duplication history accordingly.

We present algorithms for determining the history of duplications in rooted and unrooted gene family trees automatically in Section 2. We implemented these algorithms and tested them on gene family trees published in the molecular evolution literature [10, 23, 24]. As summarized in Section 3, the hypotheses for rooted trees generated by our program are consistent with the verbal assessments made by the biologists who constructed and evaluated the trees. For unrooted trees and nodes with low bootstrap values, our program generates and scores all alternate hypotheses, providing an exploratory analysis tool for considering alternative interpretations. Since the number of alternatives for even a small unrooted tree is quite large and can increase dramatically when bootstrap support is weak, automatic generation of alternate hypotheses will be necessary, while scoring those hypotheses reduces the search time for the user. In addition, an explicit statement of all hypotheses helps mitigate any biased expectations of the data the user might have.

Our tool is also a step towards the automated analysis of duplications in large genomic data sets. 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 [29], Wolfe and Shields identified 446 duplicated genes in 55 putative duplicated regions. This data set is an order of magnitude larger than the gene duplication studies currently being carried out "by hand".

Related Work. In contrast to the the interest in duplication in the biology literature, most computational work on genome evolution has focussed on reconstructing the history of speciation rather than the evolution of genome structure in a single lineage. Sankoff [25] introduced the idea of using the minimum set of rearrangements necessary to transform one genome to another as a measure of the relatedness of the two species. This led to a series of algorithmic and complexity results for computing minimum genomic distance (see [7] and the introduction to [1] for surveys). Some of these were applied to biological problems such as estimating the genomic distance between mice and humans [9] and members of the herpes virus family [8]. The set of rearrangements considered in this work includes inversions, translocations, fusions and fissions. Recently several articles have appeared that also take large scale duplication into account [2, 3, 26]. Given a genome that has sustained a single whole genome duplication followed by rearrangements, this work analyzes the minimum number of rearrangements since duplication, the structure of the preduplication genome and rates of differentiation and loss of duplicated genes.

The computational work surveyed above has been based on spatial analysis; that is, inferring the history of duplications from the positions of duplicate genes in a genome. However, unlike other genome rearrangements, temporal, as well as spatial, information can be used in analyzing the history of large scale duplications. For example, given estimates of duplication times, the likelihood that a set of duplicate genes in a putative duplicated region were copied in a single, large scale duplication can then be estimated by comparing the estimated time associated with each gene duplication. The automatic acquisition of such temporal data is the focus of the current work.

Notice that we compute a set of features on a gene family tree, based on a species tree. Another set of related work involves the relationship between gene family trees and species trees. In [5, 22], the problem was considered from the opposite perspective. Suppose that we are given a set of gene family trees involving a set of species. Can we use these input trees to produce a species tree? These papers considered, amongst other things, the problem of scoring a particular species tree with respect to a set of input trees. Their scoring function is related to the number of duplication events which the species tree induces in the gene family trees. Thus, while their goals are quite different – building species trees, rather than understanding the history of a set of duplications – we borrow some of their algorithmic techniques, as outlined in Section 2.

2 Algorithms

In this section, we present the algorithms needed for finding and dating duplications, as well as for exploring alternative hypotheses around low bootstrap edges. We consider these problems for both rooted trees and unrooted trees. The intuition for these algorithms is illustrated in Figure 3, which shows two alternate phylogenies for a hypothetical gene family, A, with two sub families, 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_2 gene has been lost due mutation or deletion or that it has not yet been sequenced. Thus, the root allows us to observe evidence of the dupli-

cation 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. We can only be sure that it took place before the evolution of birds, since there is a chicken sequence in both subfamilies.

In rooted trees, temporal order is associated with the nodes and the root defines the earliest node in the tree. As a result, a rooted tree encodes a single evolutionary hypothesis. An unrooted tree with $|E|$ edges represents up to $|E|$ different evolutionary hypotheses, one for each possible rooting. 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.

We exploit this intuition to obtain an algorithm to identify and date duplication nodes in a GFT, beginning with a discussion of rooted trees.

2.1 Rooted Trees

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 tree of life. 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 another problem, using multiple gene trees to generate the Tree of Life [5]. By using fast lca queries, M can be computed in linear time [12][1] While our goals are different, we share a key algorithmic component with this work. We refer the reader to [18] 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 the Tree of Life. 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

---

[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 inpractical. However, the state of the art of lca computation has progressed since those early papers, and there now exist lca algorithm which are very simple and very practical.

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 $U(b) = jawed\_vertebrate$.

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

Notice that we can compute U in linear time.

2.2 Unrooted Trees

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 algorithm above 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 edge 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 $M()$ 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 save value over and over, we simply store all values in a table as we compute them. Thus, once we have compute $M_{\{v,u\}}(u)$ 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 $O(n)$.

We note here that we can incorporate either NNI heuristic during the recursive computation without increasing the complexity of the algorithm.

2.3 Evaluating Alternate Hypotheses

Criteria for evaluating duplication histories are needed for comparing alternate rootings for unrooted trees. Scoring functions for duplication histories can be based on the mapping M and the assignment of duplication and speciation nodes. Each 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 development of scoring functions will require considerable experimentation in collaboration with experts in molecular evolution and is planned for future work. In the current work, we present one scoring 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 the induced Tree of Life. 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 species tree is always one.

The original motivation for this scoring function was the observation that high labels in G, labels close to M^*, tend to "trickle up" the tree. This is because M is non-decreasing along paths to the root. Given nodes x and y in G, if y is a strict 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, although

this is not guaranteed.

For every rooting of G, the root node will map to M^*. Let g be the true root of G. If G is rooted at some other node, r, then all the nodes on the path from g to r will be labeled M^*. The further the distance from r to g the greater the cost, $C(G)$.

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 one and two respectively. Each implies a single duplication. In contrast, consider the rooting shown in Figure 5. This rooting has a higher cost. 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 chicken and one copy from the second duplication was lost in all taxa except mice. 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.

2.4   Rooted Tree Reärrangements

A gene tree is inferred from sequence data with respect to some model of sequence evolution, such as parsimony or maximum likelihood. The focus of the current work is how to infer additional information on the history of duplications from a sequenced-based tree. If the sequence data does not strongly support the topology of the tree, it makes sense to consider alternate topologies for the purposes of inferring the duplication history.

As discussed in the context of the RXR example, a measure of confidence can be associated with every edge in a phylogeny. This is typically done using a resampling method called bootstrapping. Every edge, e, in a tree bipartitions the set of leaf nodes. The bootstrap value of e is the percentage of samples in which its associated bipartition was observed. 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 GFT, we consider alternate hypotheses associated with a weak edge, e by generating *Nearest Neighbor Interchanges (NNI's)* around e. This rearrangement generates alternate bipartitions for e while leaving all other bipartitions associated with the tree unchanged. Figure 6 shows a weak edge e with a pivot node, x. Removing e from the tree would partition the nodes into two sets: $(AB), (CD)$. As shown in the figure, the two NNI's associated with e involve swapping its sibling subtree, C, with the descendents of x, generating the partitions $(AC), (BD)$ and $(AD), (BC)$. The internal structure of the subtrees A, B, C and D remains intact.

An NNI will change the mapping, $M()$, resulting in a new mapping, $M'()$. In some cases, this will also change the duplication history. Figure 7(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 7(b), one rearrangement again leaves the mapping unchanged. The other rearrangement (b'') changes $M()$ and moves the duplication to the deeper node.

Changes in M associated with rearrangements can be used to evaluate alternate hypotheses, for example using cost functions like the one discussed above. Let W be the set of edges with bootstrap values below a threshold provided by the user.

Given: A GFT G, a set of weak edges W, and a TOL T.

Output: The tree G' which can be derived from G by NNI operations across edges in W such that the mapping, $M'()$, from G' to T optimizes some criterion.

Of course, the specifics of the problem to be solved vary with the criterion to be optimized. The choice of a criterion is a biological one, can include models of duplication evolution and specialized knowledge concerning the gene family.

As 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 pivot node, x, adjacent to a given weak edge, e.

Greedy NNI (GNNI) Consider a weak edge e adjacent to node x with label, $M(x)$. Let $M'(x)$ be the label of x after an 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 node, the logic being that this change may eliminate false duplications further up the tree.

Figure 7(a) represents a scenario where, presumably, 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 rearrangement in Figure 7(a) but the example in Figure 7(b) is more ambiguous. The middle tree represents a duplication before the divergence of fish followed by a loss in the fish lineage. Figure 7(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. Like the cost function, C, GNNI has implies a hidden evolutionary model: by moving duplications towards the leaves of the tree, it has the effect of selecting hypotheses that require less gene loss and fewer duplications to explain. 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 will effect the tree ultimately obtained. Since GNNI is based on the "trickle up" effect, it should be applied bottom up.

3 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. 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 presented in Section 2.3. 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 [10, 23, 24]. 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 gener-

ated results with the prose 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.* [23] used a variant in which nematodes are included in the Prostosome clade. Our standard species tree is shown in Figure 8.

3.1 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 [10]. 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:

```
Score = 0
Duplication at 15 Lower bound:
jaw Upper bound:  pro
Duplication at 14 Lower bound:
jaw Upper bound:  pro
Duplication at 6 Lower bound:
dani_reri Upper bound:  jaw
```

---

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

```
Duplication at 3 Lower bound:
xeno.laevi Upper bound:  tet
```

The score of the tree is given and time ranges for each duplication are expressed in terms of labels in the Tree of Life in Figure 8. The program finds duplications at nodes 14 and 15. 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.

3.2 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 sectioin. 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 2.4: phylogenetic corrections (e.g., Figure 7(a)) and more controversial alternate hypotheses characterized by more recent duplications and fewer gene losses (e.g., Figure 7(a)). Both types appear in the HSP70 trees shown in Figure 9. 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.

An example of NNI rearrangements for the HSP tree [10] is shown in Figure 9. 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 nine duplications and a score of three:

```
Score = 3
Duplication at 18 Lower bound:
euk
Duplication at 17 Lower bound:
euk
Duplication at 16 Lower bound:
euk
Duplication at 14 Lower bound:
```

```
        tet Upper bound:  euk
        Duplication at 12 Lower bound:
        roh Upper bound:  tet
        Duplication at 10 Lower bound:
        homo_sapie Upper bound:  roh
        Duplication at 8 Lower bound:
        amn Upper bound:  euk
        Duplication at 6 Lower bound:
        pla Upper bound:  euk
        Duplication at 4 Lower bound:
        fission yeast
```

In contrast, the topology after rearrangement (the lower tree) had three fewer duplication nodes and a score of 1. The duplications at nodes 4, 8, 10 and 12 were unaffected by the rearrangement analysis. (Nodes 8, 10 and 12 are located in the amniote subtrees and are not shown in the figure.) Duplications at nodes 5, 16 and 17 were eliminated and 14 was replaced by 13.

```
        Swapped clawed frog*hsp70 with
        8
        Swapped corn*hsp70 with
        lprn.escu3*hsc-1
        Swapped 6 with fruitfly*87c1
        Swapped 6 with 4
        Score = 1
        Duplication at 18 Lower bound:
        euk
        Duplication at 13 Lower bound:
        amn Upper bound:  tet
        Duplication at 12 Lower bound:
        roh Upper bound:  tet
        Duplication at 10 Lower bound:
        homo_sapie Upper bound:  roh
        Duplication at 8 Lower bound:
        amn Upper bound:  tet
        Duplication at 4 Lower bound:
        fission yeast Upper bound:  euk
```

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 a branching error. 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.

3.3   Unrooted Trees

We tested NOTUNG on seven unrooted trees: the TCF, CRYB and LIM families [24] , the VMAT, ANK and EGR families [23] and the PSMB family [10]. For each tree, NOTUNG computed the duplication history for every root and ranked them according to the cost function, C. We compared this ranking with the rootings favored by the authors.

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 hy-

potheses. For example, in their analysis for the TCF family tree (Figure 10), Ruvinsky and Silver [24] 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 take place prior to the amniote-amphibian separation." These conclusions are consistent with a rooting on edges (0,8), (0,4) or (0,salmon*hnf1) and no others.

For each tree, we partitioned the set of edges into plausible and implausible rootings from the verbal analysis presented by the 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. Figure 10 shows the TCF tree with rooting scores labeling each edge. In contrast to the analysis of Ruvinsky and Silver,

the edge (0,salmon*hnf1) is ranked higher than edges (0,8) and (0,4) because the scoring function favors more recent duplications. In fact, this decision should relect an evolutionary model explicitly chosen by the user.

3.4   Discussion

In this study, we analyzed every non-pathological tree in three papers [10, 23, 24]. Among these, we found no tree for which the duplication history generated or the ranking of alternate rootings was sharply at odds with the analysis of the authors of the original papers. This confirms that NOTUNG is a useful exploratory data analysis tool. The evaluation method used (C) correctly identified unlikely hypotheses, providing the user with a way to control the quantity of output to be reviewed. This approach could be improved by provide a more refined ranking. 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. With more reliable evaluation methods, NOTUNG could be adapted to the automatic analysis of large problems, such as the determination of duplication times for all paralogs in a genome. Note the 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 problematical, both in terms of increased computation time, and in terms of generating biologically reasonable heuristics.

There are several sources of uncertainty to the Gene Dating approach and data sets to which it is not easily applied. One problem is poor quality, poorly annotated or missing data. For

most species only a subset of genes have been sequenced although the percentage that remain unsequenced is shrinking. For the moment, however, most gene family trees represent partial data sets. When a paralog is not represented in a given species, we do not know if the gene has not been sequenced, if the gene was duplicated after that the divergence of that species, or if the duplication occurred and was subsequently lost due to mutation. The additional uncertainty of incomplete sequence data makes it hard to develop a reliable model of gene loss. As a result of such missing sequences, duplication nodes can be incorrectly identified. Even when the identity of the duplication node is clear, missing sequences can result in loose time bounds, since the nearest speciation node in the gene family tree may be some distance away in the tree of life.

Acknowledgements

References

[1] P. Berman and S. Hannenhalli. Fast sorting by reversal. In *CPM96*, pages 169–185, 1996.

[2] N. El-Mabrouk, D. Bryant, and D. Sankoff. Reconstructing the pre-doubling genome. In *RECOMB*, 1999.

[3] N. El-Mabrouk, J. H. Nadeau, and D. Sankoff. Genome halving. In Springer-Verlag, editor, *Combinatorial Pattern Matching*, pages 235–250, 1998.

[4] 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.

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

[6] 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.

[7] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[8] S. Hannenhalli, C. Chappey, E. V. Koonin, and P. A. Pevsner. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics*, 30:299 – 311, 1995.

[9] S. Hannenhalli and P. A. Pevsner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the IEEE Symposium on the Foundations of Computer Science*, pages 581–591, 1995.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] BL Maidak, GJ Olsen, N Larsen, R Overbeek, MJ McCaughey, and CR Woese. The rdp (ribosomal database project). *Nucleic Acids Res*, 25(1):109–11, 1997.

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

[19] G. Moore, T. Foote, T. Helentjaris, K. Devos, N. Kurata, and M. Gale. Was there a single ancestral cereal chromosome? *TrendsGenet*, 11(3):81–2, 1995.

[20] Ncbi taxonomy database. *http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html*.

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

[22] 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.

[23] 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.

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

[25] D. Sankoff, R. Cedergren, and Y. Abel. Genomic divergence through gene rearrangement. In *Methods in Enzymology*, volume 183, pages 428 – 438. Academic Press, 1990.

[26] C. Seoighe and K.H. Wolfe. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci U S A*, 95(8):4447–52, 1998.

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

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

[29] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.
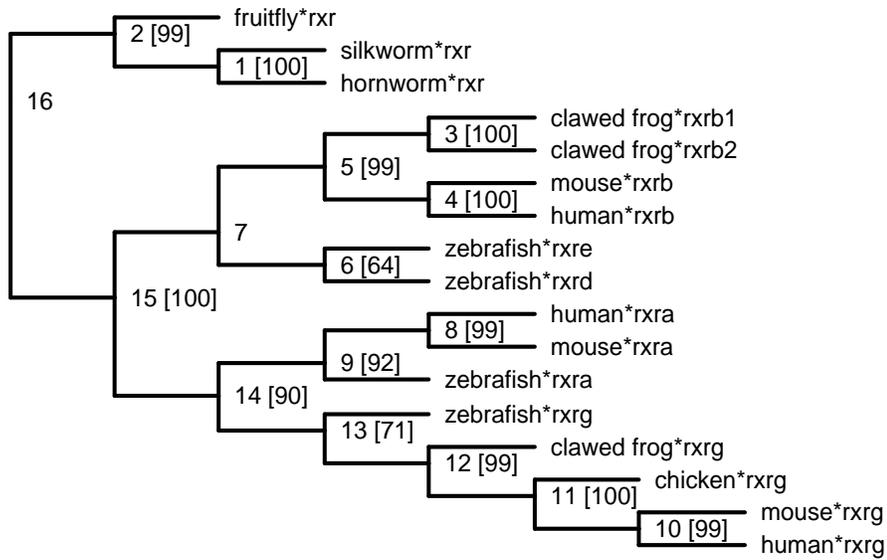
Figure 1: A rooted Neighbor Joining tree for the RXR family reproduced from [10]. Branch labels [in square brackets] represent the percentage of bootstrap samples supporting that branch. Values $\leq 50\%$ are not shown.
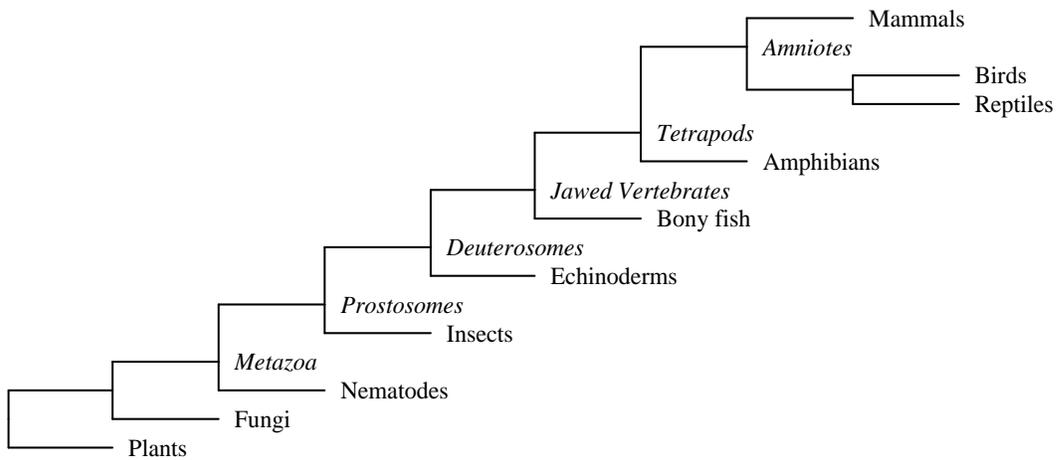

Figure 2: A species tree showing major speciation events in the eukaryote lineage. This tree was derived from the University of Arizona Tree of Life project [15] and the NCBI Taxonomy database [20]
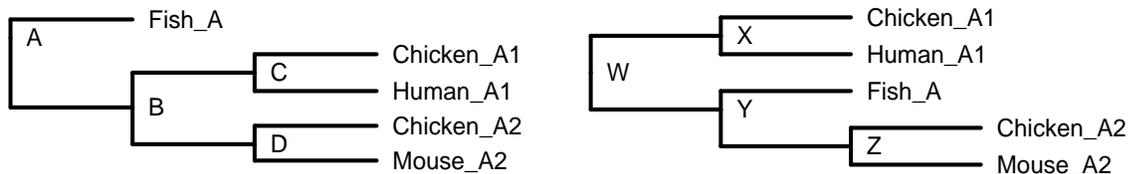

Figure 3: Two examples of rooted gene family trees representing two alternate hypotheses of the evolution of the fictional gene A family.
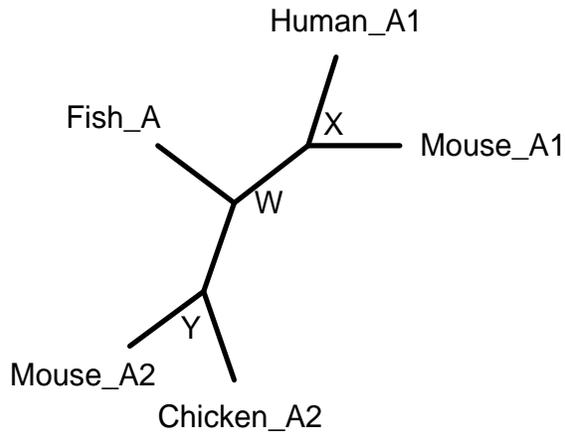
Figure 4: An example of an unrooted gene family tree for the gene A family.
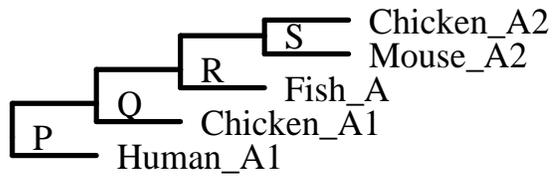
Figure 5: An unlikely rooting for the GFT for the gene A family.
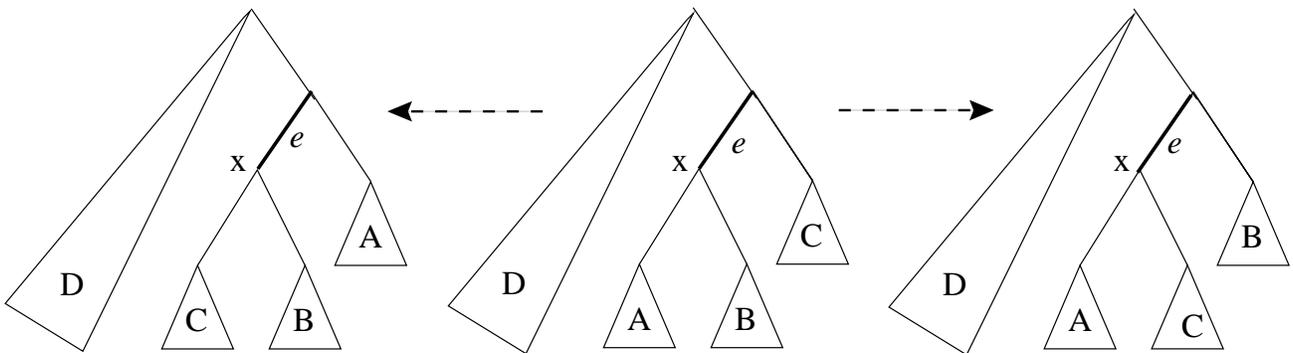
Figure 6: The three possible Nearest Neighbor Interchanges around the pivot node, x.

14
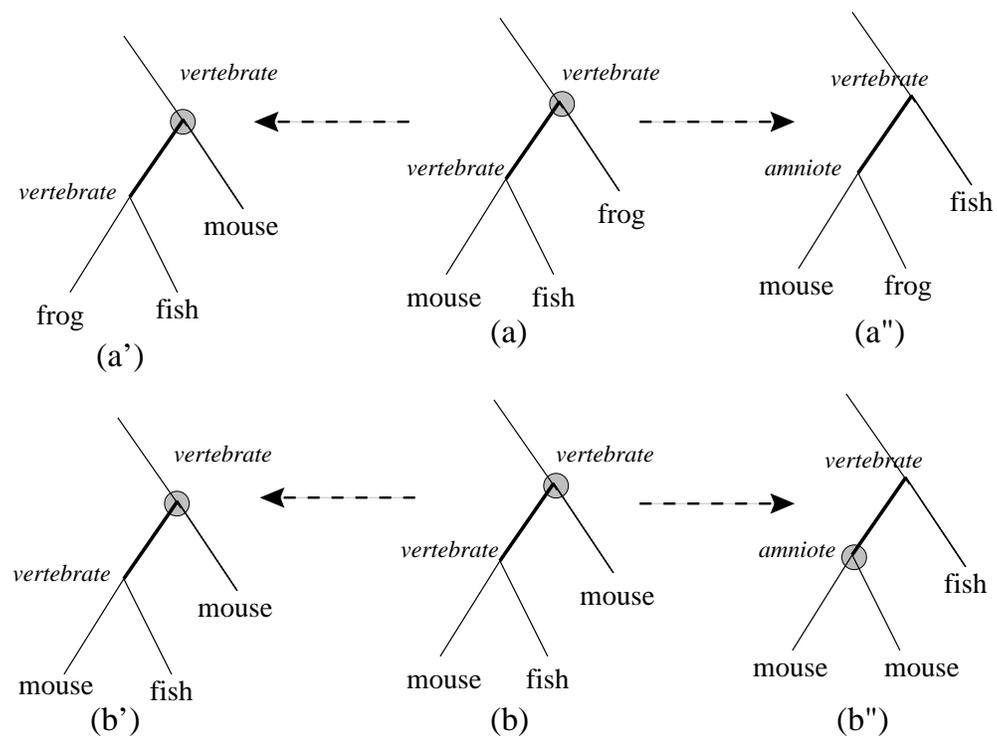
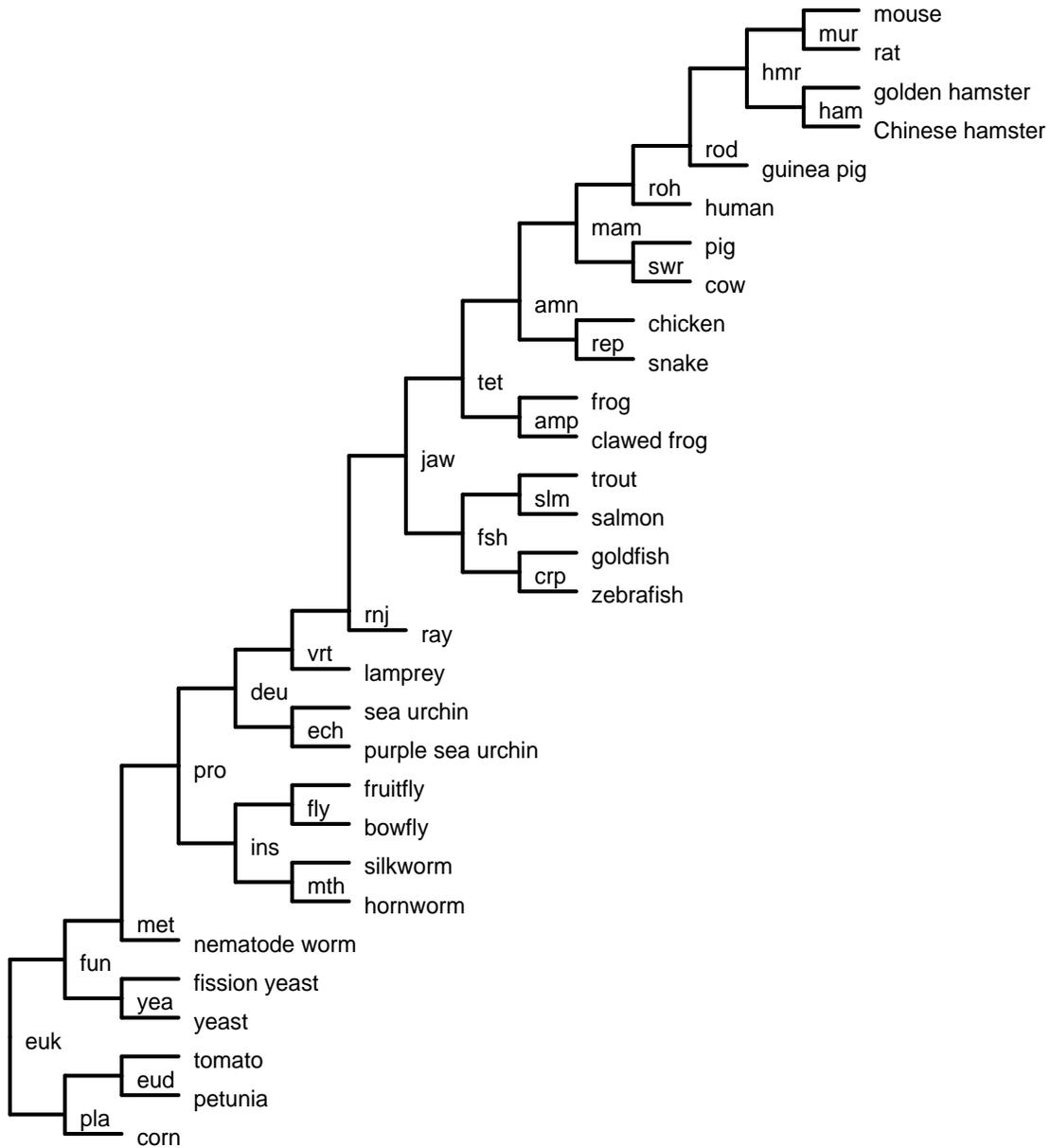Figure 7: Two tree fragments with their three NNI variants genes. Duplication nodes are shown as grey circles.

Figure 8: Standard Tree of Life used in experiments reported here. A variant was used on trees reported by Pebusque *et al.* [23]. The leaves form the union of the species represented in the datasets. This tree was constructed from information in the University of Arizona Tree of Life project [15], the NCBI Taxonomy database [20] and the Ribosomal Database Project [17, 16].
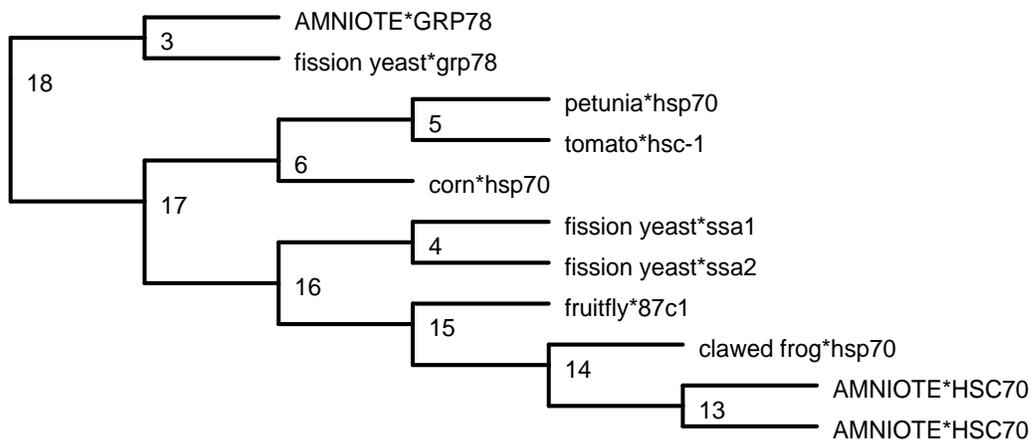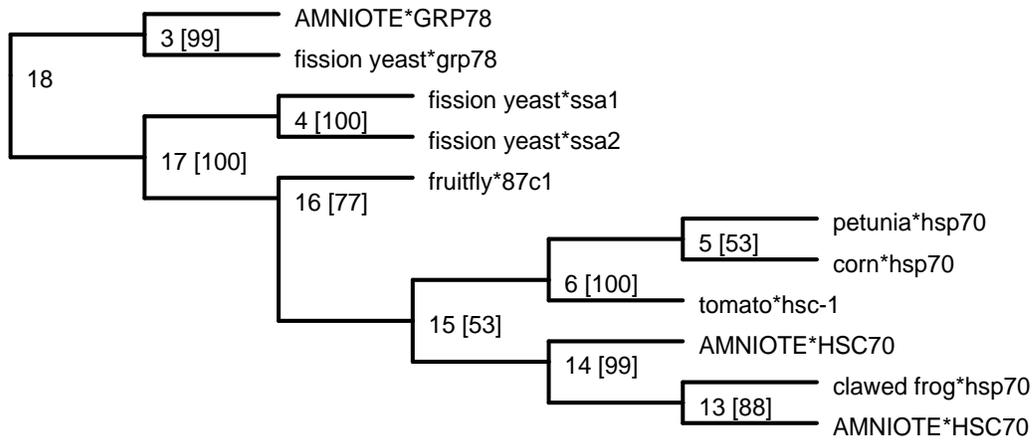
Figure 9: The HSP [10]tree before and after NNI rearrangements. The trees have been simplified by compressing clades containing only mammals and birds (AMNIOTE*GRP78m, AMNIOTE*HSP70, AMNIOTE*HSC70). No rearrangements were accepted in these clades.
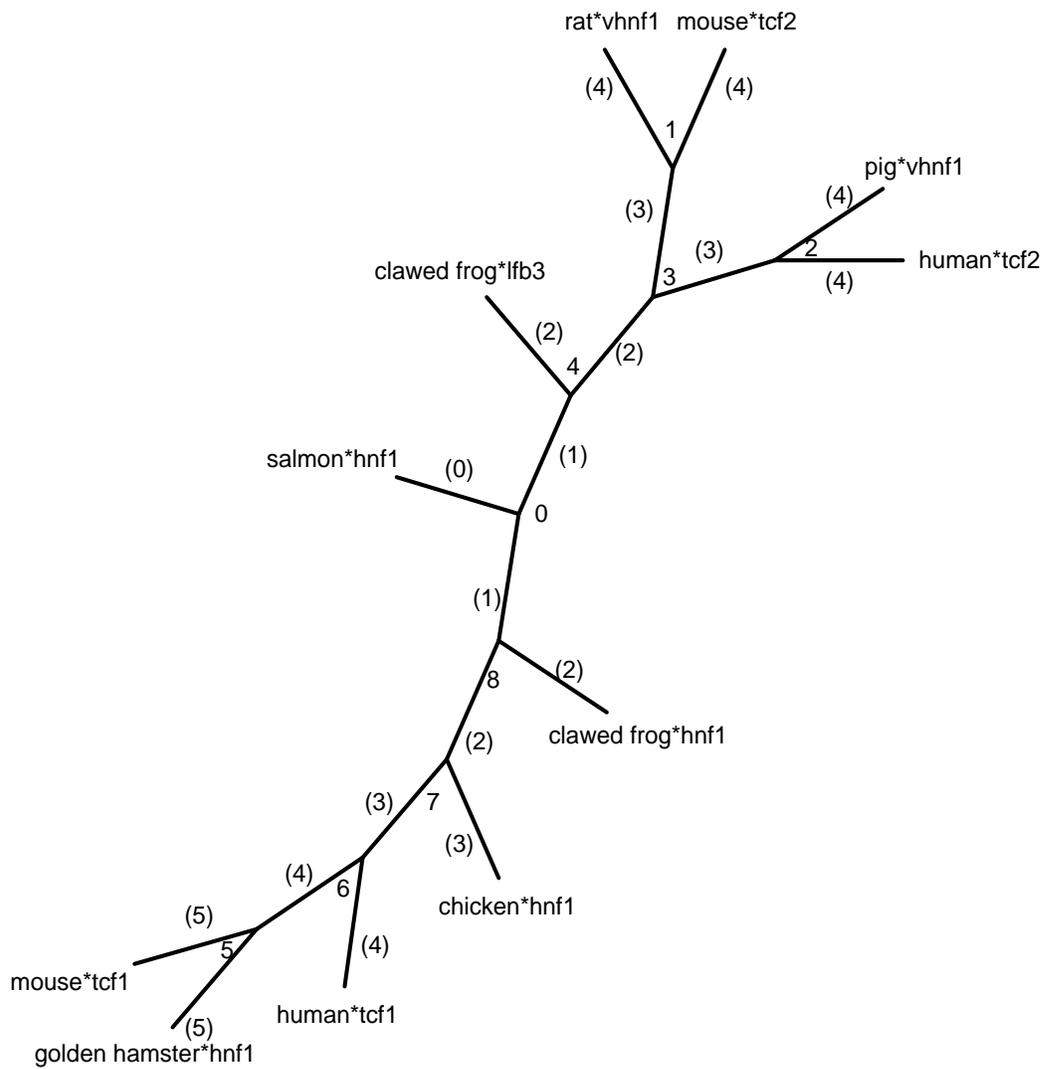
Figure 10: An unrooted tree for the TCF family [24]. Each edge, e, is labeled (in parentheses) with the cost, $C()$ of the tree rooted at e.