Notung-2.9: A Manual

Roseanna Alderson, Dave Danicic, Dannie Durand,
Aiton Goldman, Han Lai, Annette McLeod,
Maureen Stolzer, Benjamin Vernot, and Minli Xu

Date: March 3, 2020

Contents

Development of certain features, including horizontal transfer inference, was supported in part by NSF Grant DBI1262593. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 1

Introduction to Notung

Notung offers a unified framework for incorporating *event-inference parsimony* into phylogenetic tasks. This parsimony principle asserts that gene duplication, transfer and loss are rare events. Notung's functions embody the assumption that, in the absence of information from other sources, the phylogenetic hypothesis that requires the fewest number of events to explain the data is preferred.

Notung provides a graphical interface for tree manipulation and visualization and offers a command line option that can be used for automated analysis of a large number of trees. Notung can:

- reconcile a gene tree with a species tree;
- root an unrooted gene tree by minimizing events;
- rearrange rooted trees that have weakly-supported edges to minimize events;
- resolve non-binary nodes in a non-binary gene tree based on event parsimony.

Notung is used in a broad range of applications. It can assist scientists who wish to bring gene event parsimony models to bear on gene tree construction; evolutionary biologists studying the history of a gene family; and experimental biologists interested in incorporating evolutionary insights into questions of function and structure.

Notung can be used with either a duplication-loss (DL) model or a duplication-transfer-loss (DTL) event model. Several features distinguish Notung from other reconciliation programs that infer transfers. In an event model with transfers, there may be more than one most parsimonious combination of events. Notung infers all most parsimonious event histories.

Further, Notung tests all candidate event histories for temporal feasibility and screens out biologically unrealistic histories that imply a transfer that travels backwards in time. Notung reports only those event histories that are temporally feasible. Notung's transfer algorithm makes no assumptions about speciation times and does not require a species tree with branch lengths.

Notung reconciliations are detailed event histories. In addition to the reconciliation score, Notung reports specific events and the gene and species lineages in which those events occurred. Specifically, Notung reports gene duplications, and estimates upper and lower bounds on the time of duplication. Gene losses are inferred explicitly, including the location of the loss in the gene tree and the ancestral or contemporary species in which they occurred. The donor and recipient species are reported for every gene transfer.

Notung was the first reconciliation software to reconcile and root non-binary gene trees with binary species trees and binary gene trees with non-binary species trees, in addition to traditional analysis with binary gene trees and binary species trees. Notung is unique in its capability to reconcile and root binary gene trees with non-binary species trees under an event model that simultaneously minimizes transfers with duplications and losses. Another novel feature is Notung's ability to rearrange and resolve non-binary gene trees. The specific functions that Notung can perform on each combination of inputs are given in Table 1.1.

Gene Tree	**Species Tree**	**Events**	**Reconcile**	**Root**	**Rearrange**	**Resolve**
Binary	Binary	DTL	yes	yes	CL	N/A
Binary	Binary	DL	yes	yes	yes	N/A
Non-Binary	Binary	DTL	CL	no	CL	CL
Non-Binary	Binary	DL	yes	yes	yes	yes
Binary	Non-Binary	DTL	yes	yes	no	N/A
Binary	Non-Binary	DL	yes	yes	no	N/A

Table 1.1: Notung's main functions on binary and non-binary trees, with or without transfers. CL indicates the function is only available via the command-line interface.

Large-scale phylogenetic analyses can be carried out using Notung's command line interface. The phylogenomics option supports integrated analysis of all gene families drawn from a set of genomes. In addition to reporting the events inferred for individual trees, this function aggregates and summarizes results across all reconciled gene trees and all nodes/branches of the species tree. For example, it reports ancestral gene content and the combined set of events of each type associated with each node/branch in the species tree. The integrated results from all reconciled gene families can be used to infer genome-scale evolutionary trends, such as whole genome duplications, highways of lateral transfer activity, and the timing of genome expansions or contractions, relative to the species tree.

Notung utilizes novel, efficient algorithms [2, 7, 16, 25, 30] for reconstructing the history of

gene duplications, transfers and losses, for rooting gene trees based on event parsimony and for the rearrangement of weakly supported areas of gene trees based on event parsimony.

More information about Notung can be found at:

http://www.cs.cmu.edu/~durand/Notung

More information about other Durand Lab projects can be found at:

http://www.cs.cmu.edu/~durand/Lab

The graphical user interface was partially constructed using the tree visualization library provided by FORESTER (version 1.92) [32]. Cycle detection, used when inferring transfers, utilizes the jGraph library.

1.1 How to cite Notung

D. Durand, B. V. Halldorsson, B. Vernot. A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. *Journal of Computational Biology*, 13(2):320-335, 2006.

B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *Journal of Computational Biology*, 15(8):981-1006, 2008. Also appeared in Computational Systems Bioinformatics: CSB2007 Conference Proceedings, Imperial College Press: 441-452.

M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot and D. Durand. Inferring Duplications, Losses, Transfers, and Incomplete Lineage Sorting with Non-Binary Species Trees. *Bioinformatics*, 28: i409-i415, 2012.

C.A. Darby, M. Stolzer, P.J. Ropp, D. Barker and D. Durand. Xenolog Classification. *Bioinformatics*, 33(5):640-649, 2016.

H. Lai, M. Stolzer, and D. Durand. Fast Heuristics for Resolving Weakly Supported Branches Using Duplication, Transfers, and Losses. Proceedings for RECOMB International Workshop on Comparative Genomics: 298-320, 2017.

1.2 Using This Manual

This manual provides a detailed description of Notung, and gives step-by-step instructions for Notung's tasks and visualization features. It assumes familiarity with basic concepts of phylogeny reconstruction. For more information on these subjects, refer to basic textbooks, such as [10, 21]. A Glossary is provided on page 136. Additional sources are provided in the Bibliography on page 183.

The manual is organized into numbered chapters by topic. Each chapter begins with paragraphs describing the topic, followed by a list of step-by-step commands for operations associated with the topic. Figures showing the Notung graphical user interface (GUI) have been included to illustrate program displays and command results. Detailed Worked Examples are provided on page 140 to guide the user through commonly used functions and features.

Instructions for downloading Notung for various operating systems is provided in Chapter 2. A basic introduction to the Notung GUI is provided in Chapter 3. Theoretical issues related to inferring reconciliation with horizontal transfer and non-binary trees are discussed in Chapter 4. Notung's six task modes are described in Chapter 5 - Chapter 10. Chapter 11 discusses the options for changing the appearance of the tree. Detailed information about running analyses through the command-line interface, including rearrangement when inferring transfers, as well as batch processing of trees, is located in Chapter 12. More detailed information about input/output and tree file formats are given in the Appendix A. Tips regarding troubleshooting can be found in Appendix F.

Chapter 2

Downloading Notung

The Notung package can be downloaded from the Notung website as the file Notung-2.9.zip. When the file is unzipped, it will create a folder called *Notung-2.9* that includes: this manual; a folder of sample trees which contains a folder of a sample batch run; and the Notung program file, Notung-2.9.jar.

Notung is supported on Windows XP, Windows Vista and Windows 7; Mac OS X 10.5 and above; and Linux. To run Notung, Java must be installed on your computer. Notung has been tested under Java 1.5–7, but should work for newer versions of Java.

To download Notung-2.9:

Go to http://http://goby.compbio.cs.cmu.edu/Notung/download29.html

To unzip Notung-2.9.zip:

On Windows:

- If you are running Windows XP or newer, double click on Notung-2.9.zip.
- If you are running Windows 2000 or earlier, use jZip to open Notung-2.9.zip and extract its contents. If you do not have jZip, go to http://www.jzip.com and download the jZip application.

On Mac:

- Double click on Notung-2.9.zip.

On Linux:

- Run the command:

  ```
  unzip Notung-2.9.zip
  ```

If you do not know if you have Java:

- Go to http://www.java.com/en/download/help/testvm.jsp, or, in a terminal or command window, type

  ```
  java -version
  ```

  Notung requires at least Java 1.5.

To get Java (if you do not have it):

- For Windows and Linux, go to: http://java.sun.com/webapps/getjava/BrowserRedirect
- For Mac, use the Software Update application to install or update Java.

Chapter 3

Getting Started

Notung is a tool for comparing gene and species trees. Notung takes tree files as input and allows users to refine and manipulate them. The modified trees can be saved as output. The following subsections introduce basic input and output in Notung, general tree statistics, the graphical user interface, and the parameter values used in Notung's tree refinement tasks.

3.1 Gene and Species Trees

To perform its functions, Notung requires a gene tree and a species tree. The species tree must contain all the species from which genes in the gene tree were sampled. The species tree may contain additional species as well - these can be ignored. A correspondence between the leaves of the species and gene trees is determined by comparing the leaf labels in the gene and species trees: each leaf label in the gene tree must include a substring or an NHX species field that specifies the species from which the gene was sampled. Trees may be provided in Newick, NHX, or Notung format. See Appendix A - File Formats on page 126 for further information.

> *NOTE:* If your gene tree contains elements from species you do not wish to consider, the command line can be used to prune these taxa. See Chapter 12 - Command Line Options and Batch Processing on page 98 for more information.

Notung can operate on a non-binary gene tree or a non-binary species tree. However, its functions cannot be performed when both the gene tree and corresponding species tree are non-binary. In the GUI, transfer inference is limited to cases where the gene tree is binary. For a complete summary of functions that Notung can perform, see Table 1.1 on page 7.

> *NOTE:* If you are interested in using Notung to analyze non-binary trees or infer transfers, you may wish to see Chapter 4 - Theory on page 25 for more a more detailed and theoretical discussion.

Species Trees

The species tree must be rooted, with leaf nodes labeled with species names. Internal nodes may be given taxonomic labels (*e.g.*, "*tetrapoda*"), but this is not required. If the internal nodes are not labeled, Notung will assign alphanumeric labels (such as *n1, n2, etc.*). If the species tree has edge weights or branch lengths, this information will be displayed, but will be ignored for event inference. For more information on species names, see Appendix A.4 - Specifying the Species Associated with Each Gene on page 129.

The tasks that Notung performs are based on the assumption that the user has selected a species tree that is a reliable representation of the true species relationships. Using Notung with an incorrect species tree will give incorrect results. For more information on selecting an appropriate species tree, see Chapter B - Building a Species Tree on page 134.

Gene Trees

In order to perform its reconcile, rearrange and resolve functions, Notung requires a rooted gene tree. If the gene tree is not rooted, Notung can be used to root the gene tree. See Chapter 6 - Rooting Mode on page 68 for more information. The leaf nodes in the gene tree must be labeled with a unique identifier specifying the gene, as well as the species from which the gene was sampled. See Appendix A.4 - Specifying the Species Associated with Each Gene on page 129 for more information. The internal nodes may be labeled. If the internal nodes are not labeled, Notung will assign alphanumeric labels (*e.g.*, *n5, n6, etc.*).

In Rearrangement mode, Notung requires that the tree have edge weights. These are used to identify edges that are weakly supported and may be rearranged. These weights may be bootstrap values, posterior probabilities, edge lengths, or any other weighting scheme selected by the user. Several different fields in the Newick and NHX formats may be used to store edge weights. See Appendix A - File Formats on page 126 for a detailed explanation of these formats and how to indicate to Notung which field is being used for edge weights in a particular input tree.

Unrooted binary gene trees Many tree reconstruction programs represent an unrooted binary tree as a mostly binary tree, with a single trifurcation at the root. Unless a root is

selected for these trees (in Notung or another program), Notung will incorrectly treat them as rooted non-binary trees. If such a tree is actually an unrooted binary tree, failing to root it will affect Notung's diagnostics. See Chapter 6 - Rooting Mode on page 68 for more information on rooting gene trees.

3.2 The Graphical User Interface

Notung's graphical interface facilitates tree visualization and manipulation, enabling the user to inspect duplicated, transferred and lost nodes in a tree, view orthologs, paralogs, and xenologs, visualize alternate optimal trees, and color annotate genes for visual differentiation or presentation.

To run Notung:

Using the graphical user interface on Windows or Mac:

- Unzip the downloaded file.
- Double-click on the file Notung-2.9.jar.

Using the graphical user interface on Linux:

- Unzip the downloaded file and move into the Notung directory.
- Enter in the command line:

  ```
  java -jar Notung-2.9.jar.
  ```

  - OR -

- Enter in the command line:

  ```
  java -jar "PATH_TO_NOTUNG"
  ```

In addition, Notung can perform many of its operations from the command line without launching the GUI. (See Chapter 12 - Command Line Options and Batch Processing on page 98 for a description of the command line interface.)

When Notung is first launched, the program window will be blank. Figure 3.1(a) and Figure 3.1(b) show Notung's graphical interface once a gene tree and species tree have been opened. Notung's graphical user interface has the following components:

Tree panel: The tree that is currently selected appears in the tree panel. Trees are rendered with the root at left and leaf nodes at right. Nodes are denoted by small blue squares in the tree. Edge weights and leaf node names appear in the tree by default. Notung fits the whole tree in the tree panel by default. The size of the tree and tree labels can be modified using the Zoom and Fonts menus, respectively. See Chapter 11 - Changing the Appearance of the Tree Panel on page 93.

Although multiple trees can be open in Notung at once, Notung operates on only one tree at a time. To facilitate working with many trees, Notung marks each open tree with a tab at the top of the tree panel. Clicking on a tab selects the corresponding tree. Tabs are labeled with the file name and special icons to identify them as a gene or species tree - a DNA helix for gene trees, and a cartoon of the evolution of humankind for species trees (see Figure 3.2).

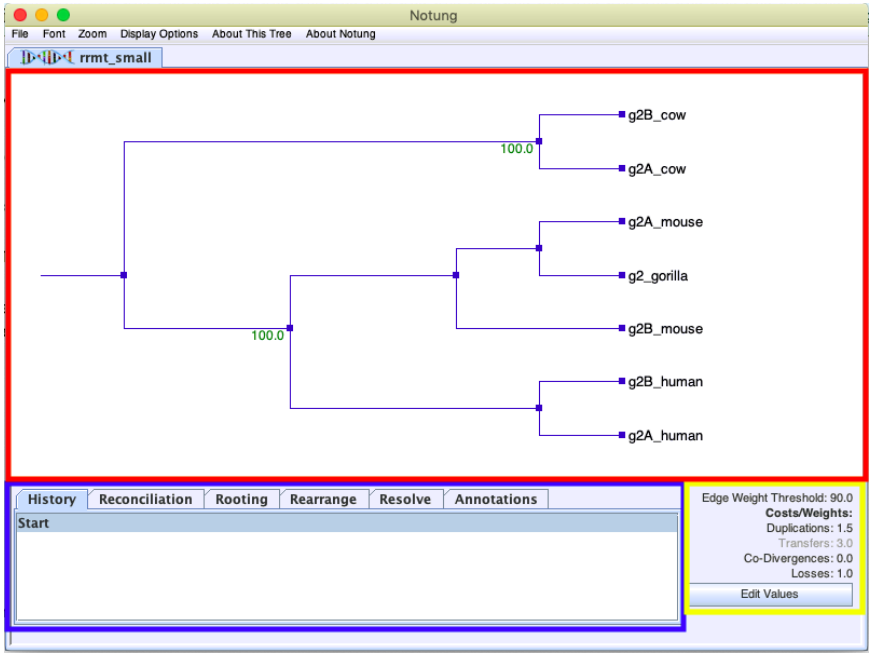
Task panel: Operations on the tree are performed in the task panel (highlighted in blue in Figure 3.1). Tabs at the top of the task panel correspond to the various tasks that Notung can perform. Clicking on a tab puts Notung in the corresponding task mode, revealing the buttons that control tasks specific to that mode. If a gene tree is selected, six modes are available: History, Reconciliation, Rooting, Rearrange, Resolve, and Annotations. Only the History and Annotations modes can be used when a species tree is selected.

Parameter values: When a gene tree is selected, a box displaying the Edge Weight Threshold and Costs/Weights for Duplications, Transfers, Co-Divergences, and Losses appears in the bottom-right corner of the program window. These values can be changed by clicking the "**Edit Values**" button directly below them. A value for transfers can only be edited if the "Infer Transfers" option is selected. Note that when a species tree is selected, the program window will not display the parameter values.

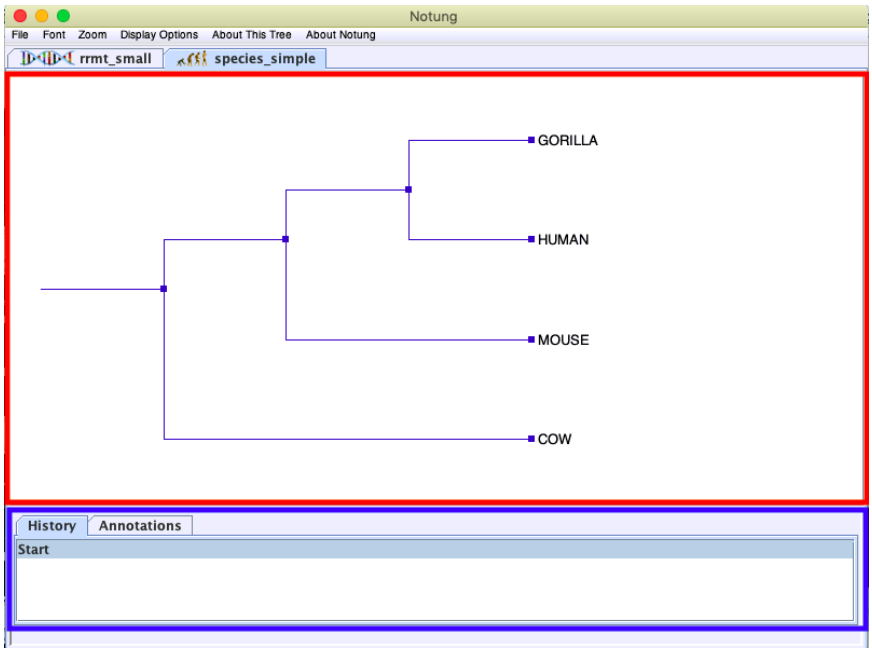
3.3 File Menu and Opening and Saving Trees

Notung can read and save tree files in Newick, NHX, and Notung file formats. NHX and Notung file formats are extensions of Newick; see Appendix A - File Formats on page 126 for details. Notung can also save the image in the tree panel as a Portable Network Graphic (PNG) file.

To open trees:



(a)



(b)

Figure 3.1: Notung's graphical user interface displaying (a) a gene tree, and (b) a species tree. The tree panel is highlighted in red, the task panel in blue, and the parameters panel in yellow. Only the tree panel and the task panel are applicable to species trees.

1. Click "**File → Open Gene Tree**" or "**File → Open Species Tree.**"

Figure 3.2: Tree tabs for a gene tree (left) and a species tree (right)

2. In the Open dialog box, select a tree file and click "**Open**."

   NOTE: Notung cannot distinguish gene trees from species trees automatically. If a gene tree is opened as a species tree, or a species tree is opened as a gene tree, reconciliation will produce incorrect results.

To save trees:

1. Click "**File → Save As.**"
2. In the drop-down menu, "Files of Type," select one of the following formats:
   - Notung File Format
   - Newick File Format
   - NHX File Format
3. Click "**Save**."

   NOTE: The default format for saving trees is the Notung File Format. If you have modified the tree in Notung and wish to reopen this tree in Notung, it may be best to save the tree in Notung format. If you wish to reopen the modified tree in another tree program, Newick format may be a better option. See Appendix A - File Formats on page 126 for more information.

To view text formatted trees in a dialog box:

1. Click "**File → View Tree in Text Format.**"
2. In the drop-down menu, select one of the following formats:
   - Tree in Notung Format
   - Tree in Newick Format
   - Tree in NHX Format

   To copy this information, click the "**Copy to clipboard**" button. This text can then be pasted in any text editor.
3. When finished reviewing this information, close this window to continue using Notung.

To save the current view of a tree as a PNG file:

- Click "**File → Save Current View as Image (PNG)**."

  NOTE: This option saves only the image currently visible in the tree panel. If you have zoomed in on a tree, the PNG will save only the section in view.

To save an image of the whole tree as a PNG file:

- Click "**File → Save Whole Tree as Image (PNG)**."

  NOTE: This option saves a "pretty print" version of the entire tree. Currently, display options set in Notung will not affect the output of this tree. More options for saving tree images are available via the command line, as discussed in Section 12.5 - Saving PNG Images of Trees

To print an image of a tree:

1. Click "**File → Print Current View**."
2. The print dialog box will appear. Change the settings as necessary and click "**Print**."

   NOTE: For most printers the default page layout will be portrait; however, the landscape layout is usually preferred for printing trees from Notung. You may wish to change your printer settings before printing.

3. A red rectangle will appear in the tree panel. Only the view inside this rectangle will be printed.
4. To proceed with printing, click "**Print**."
5. If you wish to change the printer's settings or the size of the tree, click "**Cancel**." The red rectangle will disappear and the appearance of the tree can be manipulated.

   NOTE: Printing a view of the tree that shows exactly what you want may be difficult as it may be necessary to change both the printer's settings (*i.e.* page layout, margins, *etc.*) and the appearance of the tree so that the desired print area fits within the red rectangle. See Chapter 11.2 - Zoom on page 95 for more information on zooming in and out of the tree. It may be easier to obtain the desired view by first saving the tree as a PNG image, and then editing and printing that image using another program.

To reload a tree:

- Click "**File** → **Reload from file**."

  NOTE: If the tree has been modified, a dialog box will be displayed. The dialog box will offer you one of the three following options: "Save tree"; "Reload tree without saving"; "Cancel reload".

To export color annotations to a file:

1. Click "**File** → **Export Annotations**."
2. Provide a file name and click "**Save**."

   NOTE: Exported annotations can be imported into other trees, or loaded on the command line using the option `--annotationfile`. For more information about color annotations, see Chapter 10 - Annotations on page 87.

To import color annotations from a file:

1. Click "**File** → **Import Annotations**."
2. Select the desired annotations file and click "**Open**."

   NOTE: Annotations can be imported from previously exported annotations files. Additionally, selecting a Notung format tree which contains annotations will import annotations from that tree. Annotations can also be loaded via the command line using the option `--annotationfile`. For more information about color annotations, see Chapter 10 - Annotations on page 87.

To close trees:

1. Select the tree to close.
2. Click "**File** → **Close**."

To quit Notung:

- Click "**File** → **Exit**."

3.4 General Tree Statistics

Notung compiles information on tree characteristics, such as height, number of leaves, number of nodes, *etc.* Notung reports this information is reported in the general tree statistics pop-up box under the "About This Tree" menu. The properties examined depend on whether the given tree is a gene tree or a species tree, and whether the gene tree has been reconciled or not. A description of the possible information displayed is described below.

For all trees

Statistics for: the file name for the given tree.

Total nodes: the total number of nodes.

Internal nodes: the total number of internal nodes (**Total nodes** minus **Leaf nodes**).

Leaf nodes: the total number of leaves.

Polytomies: the total number of polytomies in the tree. This number will be zero if the tree is binary.

Size of largest polytomy: the number of children of the largest polytomy in the tree. This number will be zero if the tree is binary.

Height: the maximum path length from a leaf node to the root.

Figure 3.3 shows an example of the tree statistics provided for a species tree.

For gene trees, but not for species trees

Edge Weight Range: the range of edge weights in the gene tree in the form, [minimum edge weight, maximum edge weight].

For reconciled gene trees

If the tree has been reconciled and the current solution is feasible, additional information on the number of inferred events is provided under the heading *Reconciliation Information*:

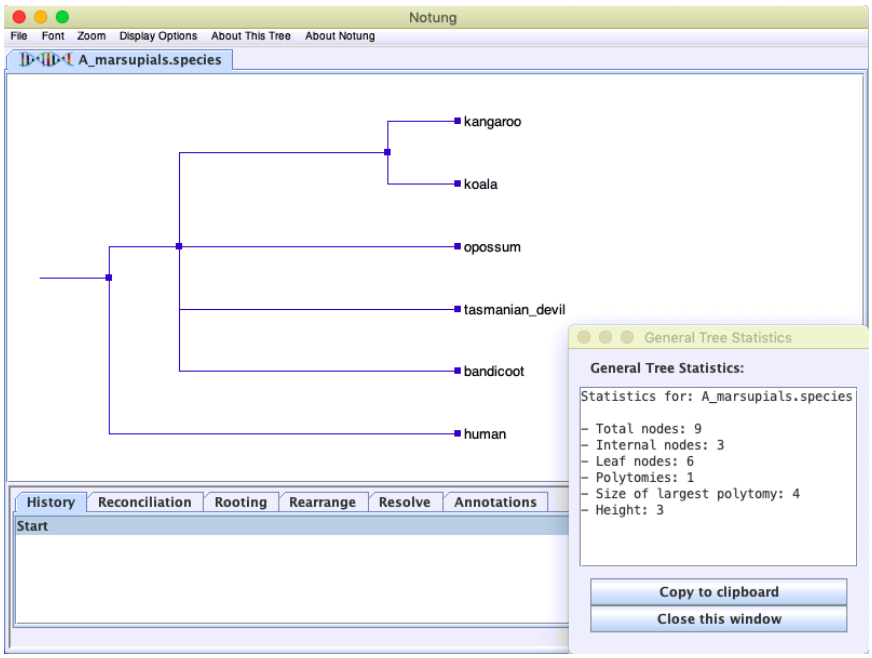


Figure 3.3: General tree statistics for a species tree.

Duplications: the total number of duplications in the reconciled gene tree.

Transfers: the total number of transfers in the reconciled gene tree. This number will be 0 if only duplications and losses are optimized, or if no transfers are inferred.

Co-Divergences: the number of co-divergences in the reconciled gene tree. This number will be zero if the associated species tree is binary or there are no co-divergences. See Chapter 4.2 - Non-Binary Trees on page 29 for more information on co-divergences.

Losses: the total number of losses in the reconciled gene tree.

This is followed by statistics on the topology of the reconciled tree (number of leaf nodes, number of internal nodes, *etc.*) with and without losses, and then by statistics on the topology of the associated species tree. If the tree has been reconciled, but there are no feasible solutions, a message stating such will be reported and include the file name for the associated gene or species tree along with the current event costs.

Figure 3.4 shows an example of the tree statistics displayed for a reconciled gene tree.

To get general statistics for a tree:

- Click "**About This Tree → General Tree Statistics**."

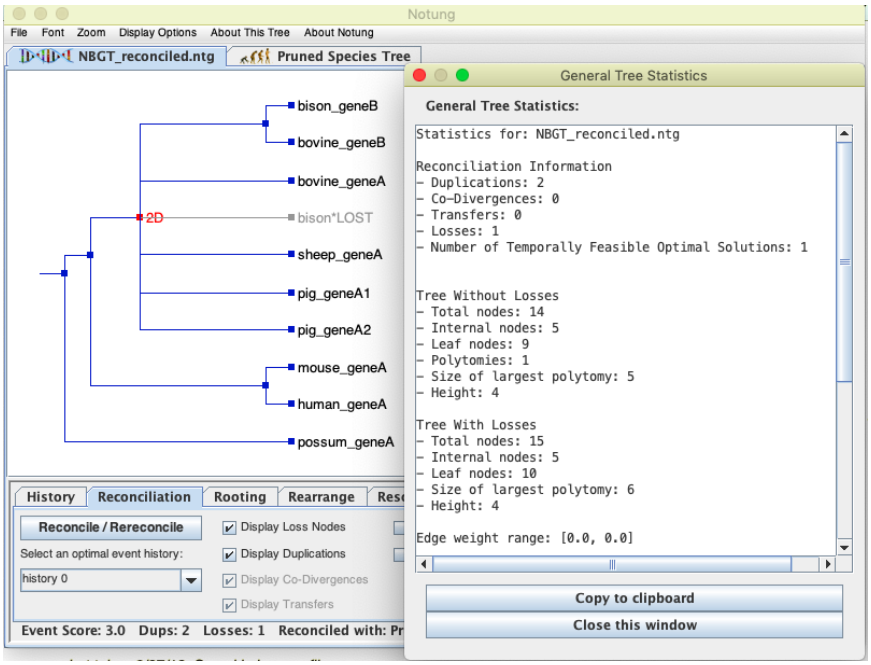


Figure 3.4: General tree statistics for a reconciled gene tree.

A window will appear containing information on the tree's characteristics, as described above. To copy this information into your favorite text editor, click the "Copy to Clipboard" button, and paste in the text editor.

NOTE: Information on duplications, transfers, and losses can also be gathered through the "About This Tree Menu" with Event Summary. For more information, see Chapter 12.2 - Duplication Bounds and Loss Information on page 104.

3.5 Parameter Values

The parameter values used in Notung — the Edge Weight Threshold, Duplication Cost, Transfer Cost, Co-Divergence Cost, and Loss Cost — can be specified by the user. These values influence the results produced by Notung's tasks.

Notung uses a Duplication/Transfer/Loss (DTL) Score to score reconciled trees and evaluate alternate hypotheses. The **DTL Score** is defined to be: $c_L L + c_D D + c_C C + c_T T$ where L is the number of losses, D is the number of duplications, T is the number of transfers, and C is the number of co-divergences implied by the current reconciliation. The loss cost, c_L, duplication cost, c_D, transfer cost c_T, and co-divergence cost, c_C reflect the relative importance of losses,

duplications, transfers and co-divergences in scoring the tree. The cost of co-divergences is only relevant when reconciling a gene tree with a *non-binary* species tree (see Chapter 4.2 - Non-Binary Trees on page 29). Likewise, the cost of transfers is only relevant when including transfers in the optimization. The default cost values are 1.0 for losses, 1.5 for duplications, 3.0 for transfers, and no cost for co-divergences, but these values can be changed by the user. Notung displays the DTL Score of a reconciled tree, as well as the number of losses, duplications, transfers and co-divergences, in the bottom-left corner of the program window (see Figure 3.5).

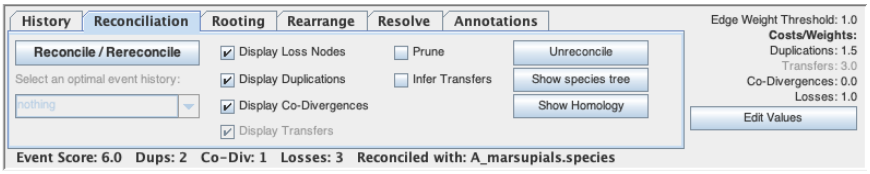


Figure 3.5: If the gene tree has been reconciled, the DTL Score, the number of duplications, transfers, co-divergences and losses, the number of optimal solutions, and the species tree used to reconcile it appear at the bottom of the program window.

The Edge Weight Threshold is a parameter used to define the set of strong edges in the gene tree. In Rearrange mode, edges weighted below the Edge Weight Threshold are considered weak and may be rearranged (for more information about rearrangement, see Chapter 7 - Rearrange Mode on page 73). Edges with no weight specified are assigned an edge weight of zero, and are considered to be weak. The default threshold is 90% of the highest edge weight in the gene tree file. If no edge weights are found, the threshold is set to one. The user may change this cutoff if a different threshold is desired for the current data set.

> *NOTE:* For some sources of edge weights, such as bootstrap values, setting the threshold to a percentage of the highest edge weight works well. For other sources, such as branch lengths, where a single very large value could cause all other edges in the tree to be weak, it may be better to set the threshold with a fixed, minimum value.

To change the parameter values:

1. Click the "**Edit Values**" button. A dialog box appears.
2. Enter the appropriate values in the text field, and then click "**Apply Changes.**"

   > *NOTE:* This will change the value settings only for the gene tree that is currently selected. Also, each history state saves the parameter values used at that state; when moving through the history, parameter values may change depending on the state and tree viewed. For more information on history

states, see Chapter 9 - History on page 85.

NOTE: Transfer costs can only be edited if the "Infer Transfers" option is selected.

Chapter 4

Theory

If you plan to infer transfers or to analyze either non-binary gene trees or non-binary species trees using Notung, this chapter is recommended. If you will be working solely with binary trees and a duplication-loss event model, you may skip ahead to the chapters describing the specific tasks you wish to perform.

Reconciliation relies on the observation that discordance between gene and species trees is evidence that genes diverged through processes other than speciation, including gene duplication, gene loss, and horizontal gene transfer. Reconciliation establishes a correspondence between the evolutionary histories of genes and species. This correspondence, called a mapping, is constructed between nodes in the trees, revealing which ancestral entities coexisted and where gene and species evolution did not proceed in a coordinated manner. The events that best explain such incongruence are then inferred. In a parsimony framework, these are defined as the events that minimize the Event Score, a weighted sum:

$$\pi = c_D D + c_L L + c_T T + \ldots \tag{4.1}$$

The reconciliation of a gene tree with a species tree results in an annotated gene tree, in which every internal node is annotated with the species that contained the gene and the event that caused the divergence. In addition, edges in the gene tree are annotated with the genes lost, labeled with the species in which the loss occurred. In a parsimony framework, the event inference problem can be stated formally as follows:

The Event Inference Reconciliation Problem

Input: A rooted species tree; a rooted gene tree; the leaf mapping, indicating from which species the genes were sampled; an event model.

Output: All reconciliations that minimize the weighted sum of the inferred events.

Notung's default mode is reconciliation of a binary gene tree and a binary species tree under a duplication-loss (DL) event model. It minimizes the DL score,

$$\pi = c_D D + c_L L. \tag{4.2}$$

Algorithms designed for Notung utilize this DL Score to reduce uncertainty in gene trees by highlighting promising roots [2] and rearranging weakly supported areas [7]. See Chapter 6 - Rooting Mode on page 68 and See Chapter 7 - Rearrange Mode on page 73 for more information.

Alternatively, users can select a duplication-transfer-loss (DTL) model, which minimizes a DTL Score,

$$\pi = c_D D + c_L L + c_T T. \tag{4.3}$$

This model is discussed further in Section 4.1.

> *NOTE:* Event costs are supplied by the user. Higher costs for transfers, compared to duplication and loss, should be used in situations where duplications are thought to be the dominant event, while the opposite should be utilized in cases where transfers are the dominant event (e.g., in bacteria).

The user may also opt to consider either a non-binary species tree or a non-binary gene tree (but not both at the same time). These non-binary tree functions are based on algorithms we developed for Notung [25, 30] and are discussed further in Section 4.2.

> *NOTE:* Notung can only infer transfers for a binary gene tree. For a complete listing of the functions that Notung is able to perform on binary and non-binary trees, see Table 1.1 on page 7. For more information on all the algorithms implemented in Notung, please see our work listed in Chapter 1.1 - How to Cite Notung.

4.1 Transfers

Horizontal gene transfer, the transmission of genetic material from an organism in one species to the genome of an organism in another species, is a common phenomenon in prokaryotes. While the extent and importance of horizontal transfer in eukaryotes is less well-understood, a growing body of evidence for HGT in eukaryotes (e.g., [1, 31]) and duplications in prokaryotes (e.g., [24]) calls for models that capture both events [9]. Most event-inference reconciliation algorithms consider either gene duplication or HGT [6, 19, 20], but not both. Our algorithm

for binary species trees is inspired by the Duplication-Transfer (DT) models of Tofigh et al. [29], but differs in that it includes losses in the optimization criterion. In our model, the event history that best explains the disagreement between gene and species tree is the set of duplications, transfers and losses that minimizes the DTL Score.

In Notung, transfers are represented by a node *and* an edge. If a transfer is the cause of a divergence (internal node) in the gene tree, one of the children of that divergence remains in the donor, while the other is copied into the recipient. The edge leading to the node representing the recipient's copy is annotated as a transfer edge; the parent node is assigned the transfer event.

The inclusion of transfers in the event model introduces some interesting complexities, outlined below. First, when the event model includes transfers, there can be more than one most parsimonious event history. This stands in contrast to reconciliation under the DL model, where the most parsimonious reconciliation of a given gene and species tree is unique. In addition, it is possible to construct an event history with transfers that requires traveling backward in time. Care must be taken not to infer such *temporally infeasible* reconciliations, since they obviously cannot represent the actual evolutionary history of a gene family. Both of these issues are discussed in greater detail below.

4.1.1 Multiple Solutions

More than one series of duplications, transfers and losses may result in the same pattern of incongruence because, with transfers, the gene tree is no longer confined to the structure of the species tree. When in a model with duplications, but no transfers, the ancestral species associated with a given node in the gene tree is uniquely determined by the species associated with the children of that node. However, transfers jump from one species to another, allowing gene tree nodes to map to distantly related species. In this case, the mapping from ancestral genes to ancestral species is not unique. Further, there may be two or more combinations of events with the same DTL score.

Multiple optimal reconciliations are a frequent occurrence, especially in data sets where transfer is the dominant process. To assist the user in considering these alternate hypotheses, Notung presents *all* temporally feasible, optimal solutions. In the GUI, the user may explore other optimal solutions through a point-and-click interface described in Chapter 5 - Reconciliation Mode. In the command line version, the user may ask to see several or all multiple optimal solutions by setting a flag.

Temporal Infeasibility

In order to be biologically valid, an event history that includes transfers must be temporally feasible. There is no known constructive algorithm for finding an optimal DTL-reconciliation that is temporally feasible (the problem of finding such a reconciliation is NP-complete [13]). Instead, Notung generates candidate optimal reconciliations and then checks each candidate for temporal feasibility. If a candidate reconciliation is feasible, then it is also guaranteed to be optimal; i.e., to have minimum DTL score.

Stated technically, a reconciliation is temporally feasible if it is possible to assign a temporal ordering to nodes in the species tree such that all timing constraints imposed by the species tree and the inferred transfers are satisfied. The species tree implies temporal constraints because ancestral nodes must have existed before the nodes that descended from them. Transfers imply two types of constraints. First, an inferred transfer could only have occurred if the donor and recipient species co-existed. Second, if one transfer occurred before another in the gene tree, then the donor and recipient of the first transfer must have occurred before or at the same time as the donor and recipient second transfer. These constraints are modeled using a directed graph, called a timing graph.

A history with two or more transfers could result in a situation where there is no temporal ordering that makes sense. If this is the case, the timing graph will contain a cycle. Otherwise, the graph is acyclic, meaning there exists a temporally feasible ordering of nodes.

Notung does not report infeasible solutions. We test each history generated by our algorithm to determine whether there is a feasible solution. The problem of finding the next best solution, when all the optimal solutions turn out to be impossible, has been shown to be NP-complete [13]. We currently do not have a solution to this problem. If all inferred optimally scored solutions are infeasible, Notung only provides a warning and outputs no histories.

> *NOTE:* Transfers can be inferred in polynomial time under a restricted model where only transfers between contemporaneous species are considered. This model, reviewed in [6, 15], requires estimates of speciation times, which are frequently not known. In addition, this approach may fail to recognize transfers if they involve a taxon missing from the data set [15, 19]. We make no assumptions about which donor and recipient species co-existed before transfers are inferred, except to prohibit transfers between ancestors and descendants.

4.2 Non-Binary Trees

A **non-binary**, or **multifurcating**, tree is a tree in which at least one node has more than two children. Such nodes are referred to as **polytomies**, or non-binary nodes. A polytomy can have several meanings [17]. A **hard polytomy** represents the true, simultaneous divergence of all its children. A **soft polytomy**, on the other hand, refers to the situation where the true pattern of divergence is binary, but the branching pattern is unknown. A tree with soft polytomies is sometimes referred to as an "unresolved tree." The indication is that the polytomies can be "resolved," meaning replaced with a series of binary divergences, to the true binary branching pattern.

In Notung, polytomies are represented as vertical edges with more than two children. See, for example, the polytomy in Figure 4.1.

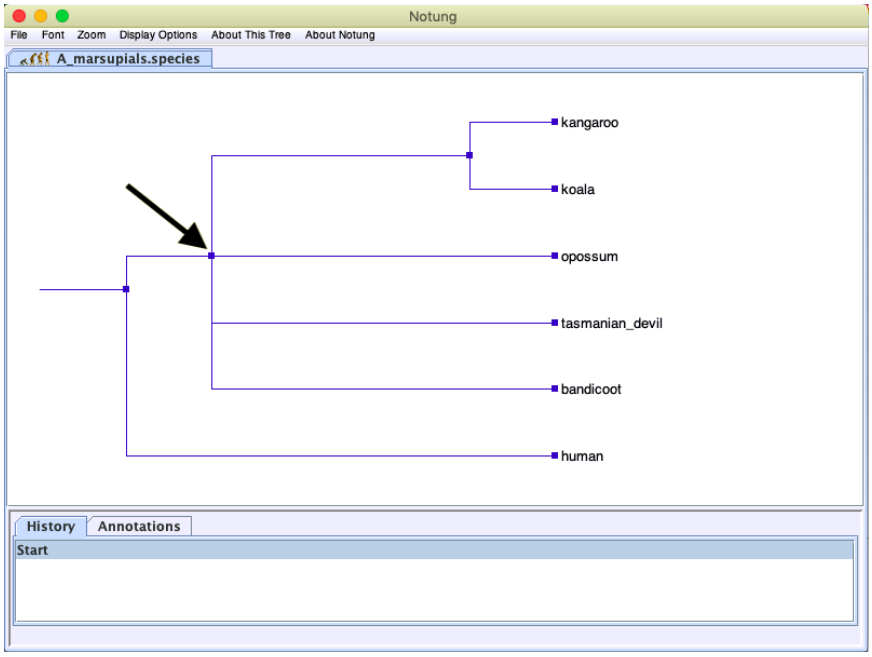


Figure 4.1: Notung displays trees as cladograms. Polytomies are drawn as vertical edges with more than two children. This tree contains only one polytomy, indicated by the arrow.

Interpreting disagreement between gene and species trees as evidence of gene duplication, transfer, and loss is widely accepted when both trees are binary. Disagreement between non-binary trees is less well-understood and there is no universally accepted approach to non-binary reconciliation. In the next sections, we briefly review current theory regarding non-binary nodes in gene and species trees and discuss how we apply these theories in Notung. First, we discuss how Notung deals with incomplete lineage sorting when reconciling binary gene trees with non-binary species trees. In the section following this, we discuss how Notung

considers the multiple, possible binary histories represented by a polytomy in a gene tree and presents the most parsimonious set of events.

4.2.1 Fitting Binary Gene Trees to Non-Binary Species Trees

Non-binary species trees may be common; for example, 64% of branch points in the NCBI Taxonomy Database [3] have three or more children.

Since a species tree represents the evolution of a population of organisms, a polytomy can be considered from two perspectives: hard or soft polytomies. Hard polytomies can result from several events, such as the isolation of subpopulations within a widespread species by sudden meteorological or geological events, or from rapid expansion of the population into open territory, resulting in reproductive isolation. Soft polytomies are frequently encountered in species trees when the signal is weak or conflicting, resulting from insufficient evidence for any particular binary branching pattern. Soft polytomies often occur if a sequence of binary divisions proceeds rapidly and the time between these events is insufficient to accumulate informative variation. These interpretations of polytomies are sometimes closely linked. The branching order of species that arose through multiple speciations in rapid successions [8, 23] is often difficult to resolve.

Incomplete Lineage Sorting

Incomplete lineage sorting (ILS) refers to discordance between gene and species trees resulting from allelic variation. Since a node in the species tree represents the evolution of a population of organisms with genetic diversity, multiple alleles may be present at the locus of interest. When lineages diverge, a different allele may fix in each lineage. The resulting gene tree will be binary and will reflect the order in which new alleles arose in the ancestral population. This pattern of divergence may not always be congruent to the pattern of divergence in the species lineage. For example, Figure 4.2 shows three different binary branching processes of a gene tree in the context of a species polytomy.

A true divergence between two genetic lineages corresponds to the point where allelic differences arose, not the time of speciation. Genetic divergence that greatly predates the time of speciation is referred to as **deep coalescence**. In Figure 4.2, for example, the divergence at x occurs much earlier than the separation of species A, B, and C, and represents deep coalescence.

The probability of incomplete lineage sorting decreases as the time between speciation events increases [14, 18, 22, 27, 28]. If branch lengths in the species tree are sufficiently long, the effect

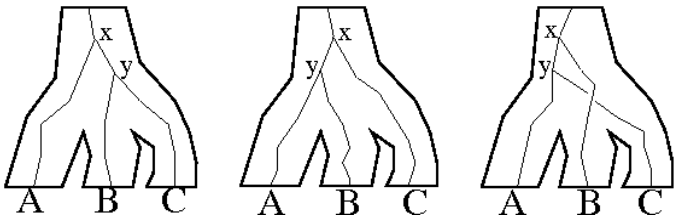


Figure 4.2: Three possible outcomes of the evolution of a single genetic locus in the context of a population. Different gene families associated with the same species polytomy may have different binary branching patterns.

of incomplete lineage sorting on discordance between gene and species trees is negligible, and does not need to be considered. However, when the species tree is non-binary, or has very short branches, incomplete lineage sorting is a plausible explanation for gene and species tree disagreement.

Inferring Events

Reconciliation focuses on incongruence that arises from processes that change the number of loci in a gene family; *i.e.*, duplication, loss, and transfer. It implicitly assumes that times between speciation events are such that genetic drift and incomplete lineage sorting may be safely excluded from consideration. This assumption breaks down when the species tree contains polytomies or very short branches. In these situations, allelic variation can survive multiple speciation events, leading to gene trees with branching patterns that differ from the species tree. Such cases are increasingly common. Whole genome sequencing data of closely related species is revealing an ever growing number of cases where ILS is active along with duplication, loss, and transfer (e.g., [1, 24, 31]), leading to calls for algorithms that model multiple evolutionary processes [5, 9]. Methods that do not consider ILS will incorrectly interpret incongruence arising from ILS as evidence of duplication or transfer.

Notung algorithmically avoids this by distinguishing between regions of the species tree where ILS is likely, and those where only gene duplication and transfer need be considered. These differences are specified using a non-binary species tree. Notung assumes that the probability of incomplete lineage sorting is negligible when a node in the species tree is binary. In this case, ILS is considered to be so rare that disagreement between the trees is interpreted as evidence only for gene duplication, transfer or loss. In contrast, if a binary gene tree node is associated with a species polytomy, by definition the trees disagree at that point. A binary divergence is not congruent with a non-binary divergence. How should events be inferred in this case?

Notung treats a non-binary node in the species tree as a hard polytomy, where disagreement can be due to the creation of a new locus, through gene duplication or transfer *or* the binary divergence of a single locus, through a deep coalescence. Both of these cases are illustrated

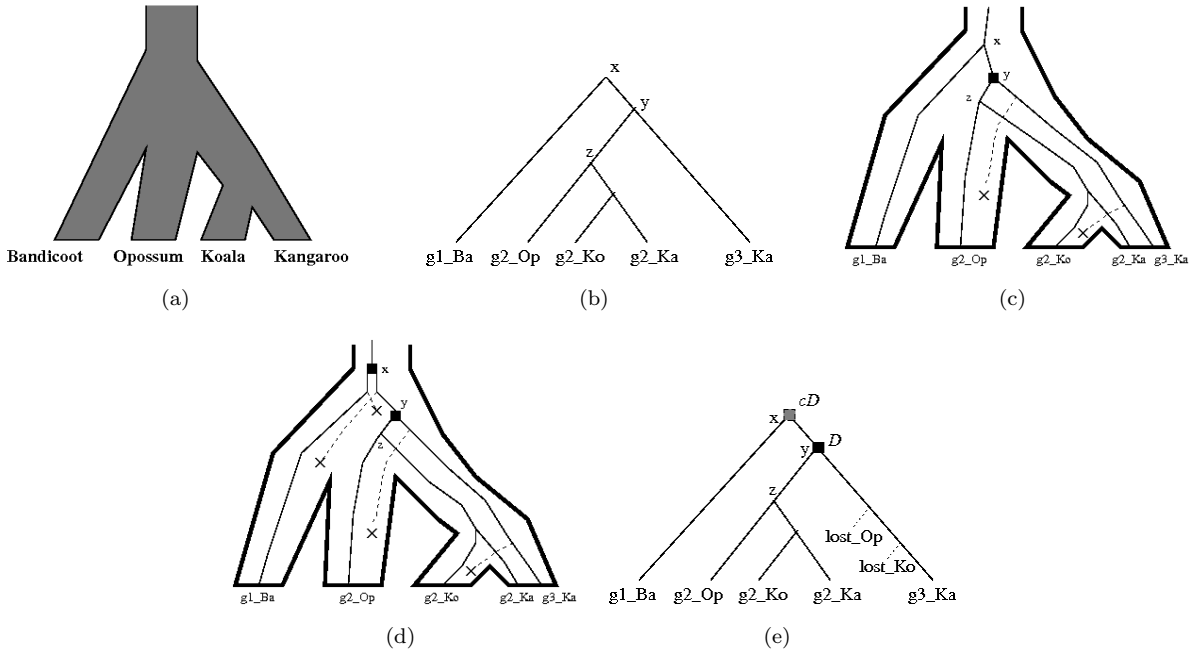


Figure 4.3: Black squares with a "D" indicate duplications. Losses are represented by dotted lines. **(a)** A marsupial species tree with a polytomy. **(b)** The phylogeny of a hypothetical gene family sampled from the same marsupial species. **(c)** Hypothesis 1: the disagreement between (a) and (b) can be explained by deep coalescence (node x), followed by gene duplication (node y). **(d)** Hypothesis 2: the disagreement between (a) and (b) can also be explained by duplication at x, followed by gene loss, followed by duplication at y. **(e)** The divergence at x is designated a co-divergence (gray square) because it is not possible to determine whether the disagreement is due to duplication or deep coalescence. The divergence at y is a duplication.

in Figure 4.3.

Our conservative approach is to assume that incongruence is due to ILS except when topological disagreement cannot be explained by ILS. This means disagreement must have resulted from the creation of a new locus through either duplication or transfer. A key aspect of our work is that, sometimes, it is not possible for ILS to explain all disagreement between the trees, *even if the gene tree is uniquely labeled.*

> *NOTE:* This model can be invoked for both non-binary species trees and for binary species trees with short branches where ILS is suspected: Even if the binary branching pattern of the species tree is known, the user can collapse edges in the species tree to indicate in which lineages ILS should be considered as an alternate hypothesis (or where incongruence should not be interpreted only as evidence of an event).

Our model allows a useful interpretation for soft polytomies. In this case, we can view the polytomy as a set of hypotheses, namely the binary resolutions of the polytomy. Since the

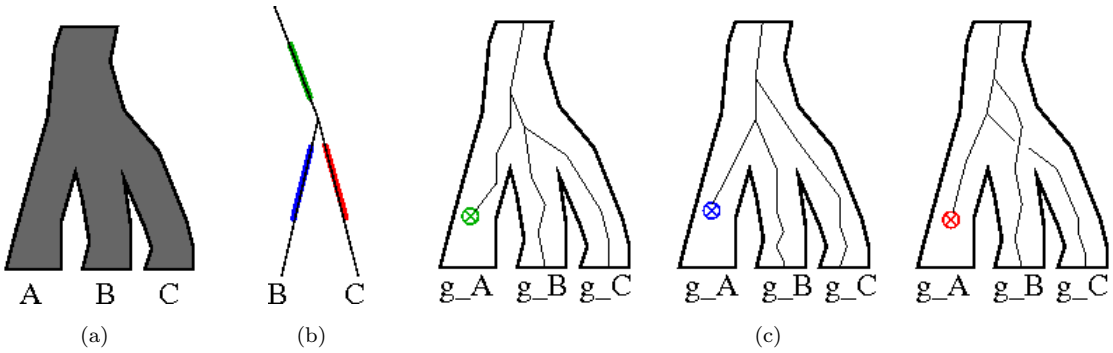


Figure 4.4: Losses associated with a polytomy in the species tree are ambiguous. **(a)** A species tree with a polytomy. **(b)** A gene tree drawn from the species in (a), with a loss in species A. **(c)** This loss can be assigned to three possible edges. Associating a loss with the green edge implies that *g_A* diverged first and was then lost. the blue edge implies that *g_A* was lost after the divergence of *g_C*; the red edge implies that *A* was lost after *g_B* diverged.

true branching pattern is unknown, it is possible that the binary branching pattern in the gene tree in fact corresponds to the true (but unknown) branching order of the species. Notung only infers an event when the gene tree does not conform to any one of the hypotheses, *i.e.*, the binary resolutions. Otherwise, the gene tree is congruent with one of the hypotheses. Notung infers that there could have been a co-divergence, and no penalty is incurred.

> *NOTE:* Notung implements novel reconciliation algorithms [25,30] for non-binary species trees that distinguishes among duplications, transfers, and co-divergences and reports them separately.

Inferring Losses

Inferring loss events is also fundamentally different when the species tree is non-binary. When both trees are binary, an inferred loss node can always be unambiguously assigned to a specific edge in the gene tree, indicating when in the history of the gene family the loss occurred. The node is labeled with the species in which the loss occurred. However, when a loss is associated with a polytomy in the species tree, it is not generally possible to assign the loss to a single edge in the gene tree. Rather, the loss can be associated with a *set* of candidate edges, each of which corresponds to an alternate hypothesis regarding when the loss occurred. The inferred loss must have occurred on one of the edges in this set, but it is not possible to determine which one. Figure 4.4 shows an example of this ambiguity when assigning a gene loss in species *A*. This loss could be associated with any of the three colored edges indicated in Figure 4.4(b). The three hypotheses resulting from the three possible ways of assigning the loss to an edge can be seen in Figure 4.4(c).

In a complex reconciliation with several losses, there may be many alternative hypotheses

(*i.e.*, reconciliations with different loss histories) to consider. Notung uses parsimony to reduce the number of candidate reconciliations. Specifically, Notung assigns each loss to a specific edge within the set of candidates, with a goal to minimize the total number of losses.

The total number of losses depends on two factors. The first is the position of the loss relative to other events in the gene tree. For example, assigning a loss to an edge *above* a duplication implies that the loss occurred before the duplication, and only one loss is inferred. However, assigning the loss to an edge *below* the duplication implies that the duplication occurred first. Thus, two losses are inferred – one for each duplicated copy. Second, in some circumstances, losses in sibling species can be more parsimoniously explained by a loss in their common ancestor. The total number of losses may be reduced by assigning losses in such a way to maximize the number of cases where multiple losses can be replaced by a single loss in an ancestral species. These two factors are not independent of one another. While assigning a loss lower in the tree will often increase the total number of losses, in some cases, these extra losses may be combined with others assigned to the same edges, thus reducing the total number of losses.

Two algorithms for inferring losses, one exact and the other a heuristic, have been implemented in Notung. The exact algorithm infers a history with the fewest losses, taking both of the above considerations into account. This algorithm is computationally intensive because all possible combinations of loss assignments must be considered. Its worst-case running time is an exponential function of the size of the largest polytomy in the species tree. In practice, the exact algorithm performs efficiently on non-binary species trees with small polytomies. However, users should be prepared for extended running times if the species tree has a polytomy with more than 12 children.

> *NOTE:* While the exact algorithm is guaranteed to return a reconciliation with a minimum number of losses, there may be more than one such optimal reconciliation; if so, Notung reports only one.

The heuristic runs significantly faster than the exact algorithm and yields the same results in many, if not most, cases. It returns only one reconciliation, which is not guaranteed to be optimal. However, in a comparison of the two methods on the 1,174 trees from TreeFam, the heuristic found an optimal solution for more than 99% of the trees. Of the seven trees where the heuristic did not find an optimal solution, in the worst case, the number of losses was overestimated by four losses from a total of 249.

Both algorithms are integrated with the algorithm to identify duplications; however, only the heuristic algorithm is used when inferring transfers. The interactive version of Notung uses the heuristic to reconcile binary gene trees with non-binary species trees. The exact algorithm is only available for duplication-loss parsimony and only in the command line version. See

Chapter 12 - Command Line Options and Batch Processing on page 98 for information about these options.

4.2.2 Fitting a Non-Binary Gene Tree to a Binary Species Tree

In a gene tree, each lineage represents a single gene and the result of any divergence is exactly two descendant sequences. Thus, in contrast to species trees, the true branching pattern in a gene tree is always binary [14], and all multifurcations are soft polytomies. For this reason, non-binary gene trees are also referred to as unresolved trees. Some phylogeny reconstruction programs output non-binary gene trees when the true binary branching process cannot be resolved. Such uncertainty often arises if binary divisions occur too rapidly to accumulate informative variation or if the data set is noisy.

Notung's approach to reconciling non-binary gene trees rests on the assumption that the children of a polytomy arose through an unknown series of binary divergences. Notung further assumes that, in the absence of other information, the best hypothesis for the true evolutionary history of the children of the polytomy is the binary branching pattern that entails the fewest duplications and losses; there may be more than one such **binary resolution** of a polytomy. The problem of reconciling non-binary gene trees reduces to finding a binary tree that agrees with the original tree everywhere except at the polytomies and has a minimal DL Score.

The general approach is as follows: A non-binary gene tree is converted into a binary gene tree by replacing each polytomy with a temporary binary resolution. This resolution is optimal under duplication-loss parsimony, when reconciled with the appropriate binary species tree. The resolution is determined by using our rearrangement algorithm [7], which constructs an optimal duplication-loss parsimony tree in polynomial time per tree. Following rearrangement, all nodes and edges not present in the original gene tree are then removed, to obtain a reconciliation of the original non-binary gene tree. As nodes and edges are removed, any duplications or losses assigned to them are reassigned to their associated polytomy.

This process is illustrated in Figure 4.5. The optimal resolution of the polytomy at node z in the gene tree in (b) with the species tree in (a) is shown in the right subtree of (c). This entails one duplication and one loss. This information is mapped onto the original gene tree (b) to obtain the reconciled, non-binary gene tree in (d). The polytomy in the original tree represents uncertainty, as reflected in the reconciliation. The reconciled polytomy in the right subtree of (d) tells us that at least one duplication and one loss occurred in the subtree rooted at z, but the exact order of these events is unknown.

Note that multiple duplications can be assigned to a polytomy in a reconciled non-binary tree. If duplications are inferred on two or more temporary nodes in the optimal binary resolution

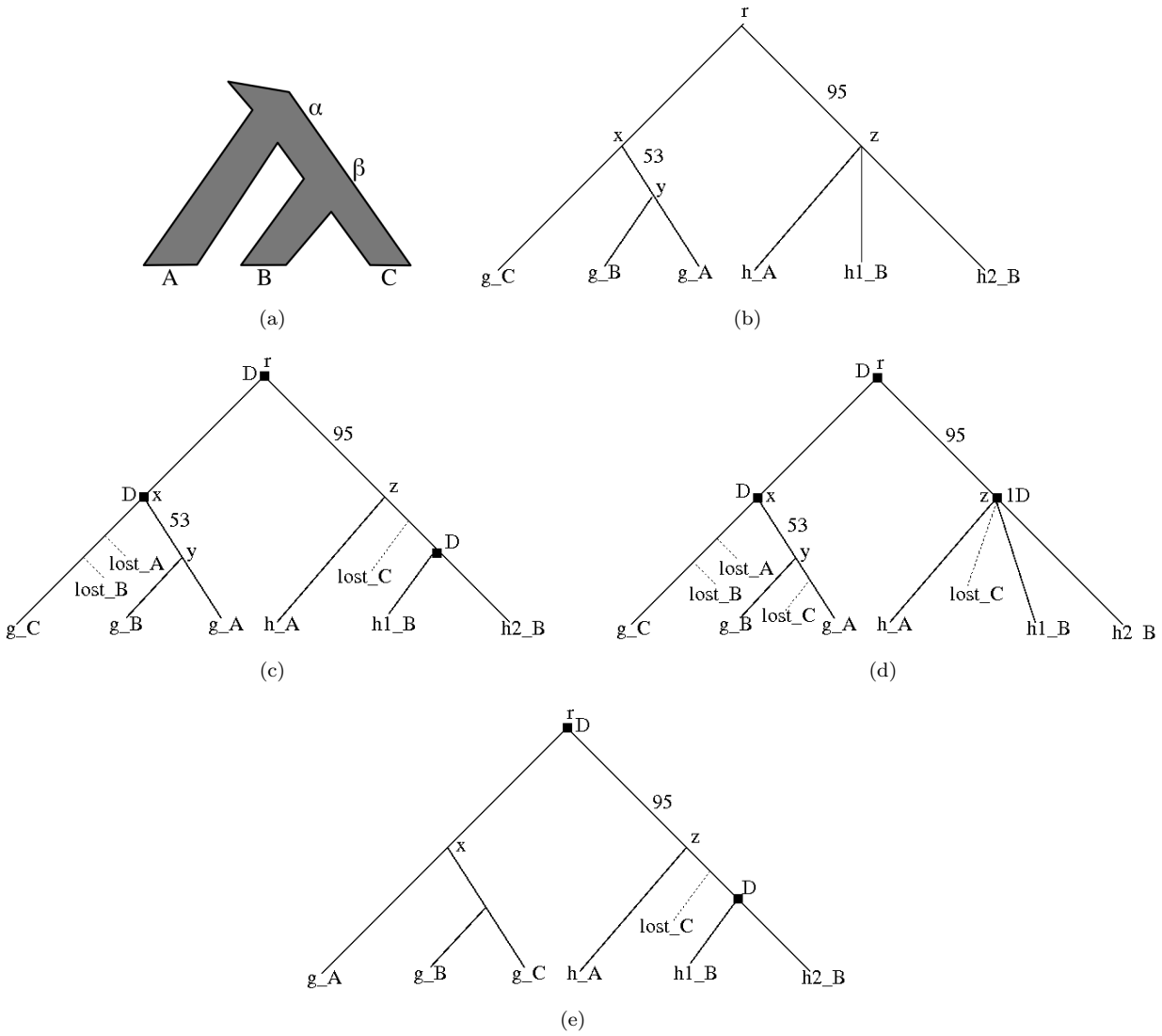


Figure 4.5: **(a)** A binary species tree. **(b)** A non-binary gene tree with genes sampled from (a). **(c)** Binary resolution of gene tree (b), yielding a binary tree with three duplications and three losses. **(d)** Gene tree (b) reconciled with species tree (a), yielding a non-binary tree with three duplications and four losses. **(e)** Gene tree (b) following rearrangement. The polytomy has been resolved and the weak edge has been rearranged to eliminate a duplication.

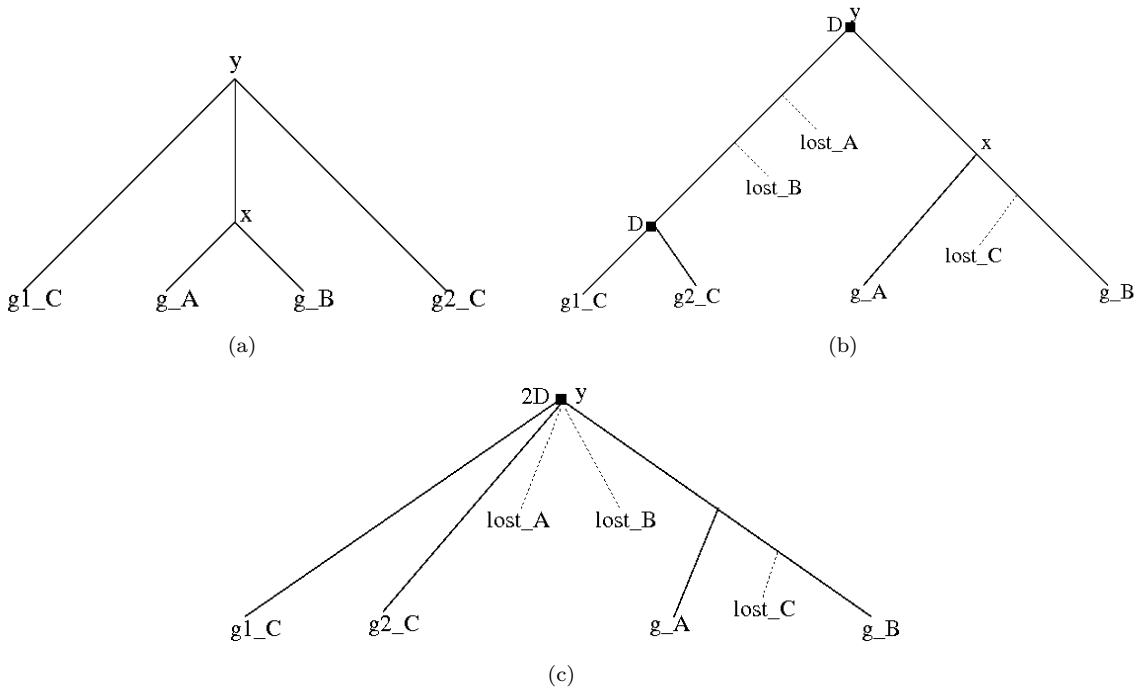


Figure 4.6: **(a)** A non-binary gene tree. **(b)** An optimal, binary resolution of gene tree (a) reconciled with species tree in Figure 4.5. **(c)** The reconciled non-binary, gene tree. The resulting tree has a polytomy with two duplications.

of a polytomy, the polytomy will be assigned multiple duplications when these nodes are removed from the tree. For example, two duplications are assigned to the polytomy in the reconciled, non-binary gene tree in Figure 4.6. This differs from standard reconciliation, where every node has at most one duplication.

Rooting a non-binary gene tree

Notung can be used to infer the root of an unrooted tree by identifying the root that requires the fewest duplications and losses. In Rooting mode, when the tree is binary, each edge is assigned a **root score**; *i.e.*, the DL Score of the tree when rooted on that edge. When the gene tree is non-binary, it is also possible to root the tree on a polytomy, as shown in Figure 4.7. Placing a polytomy at the root of the tree implies that one of the edges in the true binary resolution of the polytomy is the true root.

To calculate root scores, Notung roots the tree on each edge and polytomy in turn. For each root, the rearrangement algorithm is applied to ensure that each polytomy is replaced by an optimal binary resolution. The DL Score of the resulting tree is used as the root score for that rooting. Note that it is necessary to optimize the binary resolutions separately for each root because the DL Score depends on the location of the root. After all edges

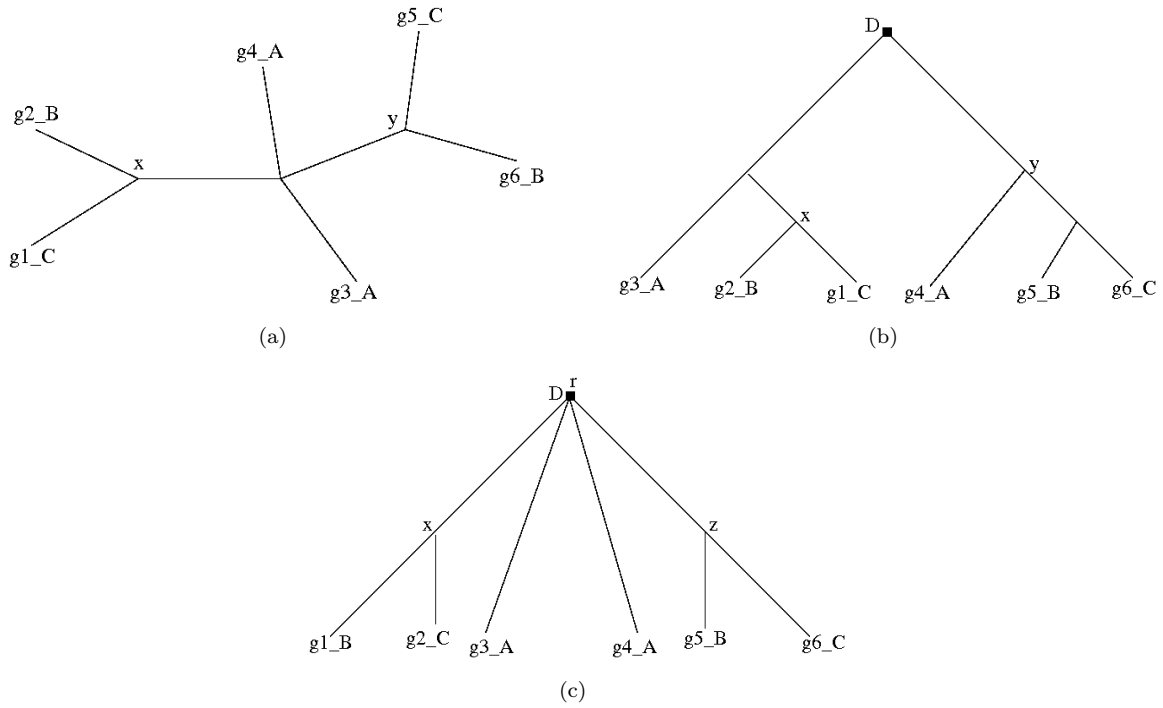


Figure 4.7: **(a)** An unrooted, non-binary gene tree. **(b)** The rooted, binary resolution of (a) with the lowest D/L Score. Rooting the tree on any other edge would entail more duplications and losses. **(c)** When reconciled with species tree in Figure 4.5(a), the polytomy in (a) is the root with minimum cost.

and polytomies have been scored, the original tree is reported to the user with edges and polytomies annotated with root scores.

Note that in Reconciliation and Rooting modes, binary resolutions are used to infer duplications and losses, but the structure of the final, output tree is unchanged. In the Rearrangement and Resolve modes, Notung uses duplication-loss parsimony to transform the non-binary input tree into a binary gene tree. Resolve mode is analogous to the reconciliation method described here, with the exception that the final step of removing the added nodes and edges is not performed. The result is a reconciled binary tree that is optimal with respect to duplication-loss parsimony. For the example in Figure 4.5, the Resolve function would return the tree in (c). As there may be more than one optimal resolution, Notung presents the different histories that result in the optimal tree. See Chapter 8 - Resolve Mode on page 81 for more information.

In Rearrangement mode, the rearrangement algorithm is applied not only to edges added to the tree in the resolution of polytomies, but to all edges with an edge weight below the edge weight threshold. The result is a reconciled, binary tree in which weak edges have been rearranged to minimize the DL Score. Figure 4.5(e) shows the rearrangement of the non-binary gene tree in (b), assuming an edge weight threshold of 90.

4.3 Practical considerations

When the species tree is binary, our transfer algorithm has $O(|V_G||V_S|^2)$ complexity when finding a single optimal reconciliation. When the species tree is non-binary, the complexity is $O(|V_G|(|V_S|+n_k 2^k)^2(h_S+k))$, where k is the size of the largest polytomy in the species tree, n_k is the number of polytomies and h_s is the height of the species tree. However, there are often multiples optimal solutions, and sometimes many, each requiring an additional $O(|V_G|)$ for binary species trees and $O(|V_G|(h_S+k))$ for non-binary species trees. Each candidate optimal reconciliation must be checked for temporal feasibility, further increasing the computational cost. The running time will increase with the number of optimal solutions. Typically, there are more optimal solutions when a large number of transfers are inferred.

While Notung will run faster if you only look at a single solution, the user should be aware that a single solution may be only one of several mathematically equivalent hypotheses. The solution arbitrarily selected by Notung as the first one to present to you may not represent the best hypothesis.

The complexity for algorithms with non-binary species trees will also increase with the size of the largest polytomy and the number of polytomies. In practical data analyses, k is likely to be small. Recent genome-scale analyses of ILS have focused on species trees with $k = 3$ (e.g., [8, 23]). In general, event inference will not yield informative results when the species tree is highly unresolved.

Because of the increased computational complexity of reconciliation with transfers, Notung's default settings with the DTL model differ somewhat from the DL model defaults. In these cases, the DTL default is to provide less comprehensive information in order to avoid very long running times. A more comprehensive analysis can be obtained by changing the parameter settings. Some particularly computation intensive analyses can only be run from the command line.

In the GUI, a warning message pops up in Reconcile mode if the number of candidate optimal reconciliations for gene tree exceeds 10,000. The user can choose to proceed with the reconciliation, which may require an hour or more to complete. Alternatively, the user can choose to cancel the reconciliation in the GUI, and reconcile the tree from the command line instead.

Chapter 5

Reconciliation Mode

In Reconciliation mode, Notung compares a gene tree with a species tree to infer gene duplications, transfers and/or losses. Notung will display a reconciled tree in the tree panel with the inferred events indicated on the tree. The Event Score of a reconciled tree will be displayed in the lower left corner of the screen (see Figure 5.1(b)).

Once a gene tree has been reconciled, Notung can detect homology relationships, described in Section 5.6. Notung can also determine lower and upper bounds on the time of each duplication and co-divergence, where bounds are represented in terms of internal nodes in the species tree; *i.e.*, relative to speciation events. The upper bound on the time of duplication is the most recent species in which the duplication was not present. The lower bound is the oldest species in which the duplication must have been present. This information, along with statistics on transfers and losses, can be viewed in a pop-up window by selecting "**Event Summary**" or "**Parsable Statistics**" from the "**About This Tree**" menu. Transfers, duplications and bounds in this window are identified by internal node names. For losses, each node in the species tree is listed, followed by the number of losses associated with that taxon. A table of species is also included, where rows represent each node in the species tree, and values in a row indicate the number of duplications, transfers from, transfers to, and losses associated with that taxon.

Notung can infer reconciliations with or without transfer events (*e.g.*, horizontal gene transfer), along with duplications, losses, and co-divergences. When the **Infer transfers** option in the **Reconciliation panel** is selected, Notung will allow the user to specify a transfer cost when the **Edit Values** button on the bottom right corner is clicked. Then operations including reconciliation and rooting analysis will be performed to minimize the event cost using the DTL-model, in other words, with possible transfer events taken into consideration. Please be aware that once the DTL-model is used, there are several aspects of the program the user should pay attention to, namely, (1) temporal feasibility, (2) multiple optimal solu-

tion, and (3) rearrange and resolve with transfers. These new features will be covered in the following section. For theoretical insights into reconciliation with transfers, refer to Chapter 4.

5.1 Labeling and Pruning Gene and Species Trees

Notung's reconciliation algorithms require that both trees be rooted and that each leaf in the gene tree be associated with a leaf in the species tree. For this to work, we require a way of representing the association between gene and species leaf nodes in the program's input. In Notung, the label of each leaf in the gene tree must include, as a substring, part or all of the label of the associated leaf species. This substring is called the *species tag*. Notung recognizes several different species tag formats. See Appendix A.4 - Specifying the Species Associated with Each Gene on page 129 for further information on gene labels.

The reconciliation algorithm assumes that each gene is associated with a leaf in the species tree. This is a many-to-one relationship; a species may be associated with one or more than one gene. But, what about species that are not associated with any gene? There are two strategies for dealing with this situation: These species can be removed, or *pruned*, from the tree in a preprocessing step. In this case, no penalty is incurred for observing a species that does not possess a member of the gene family. Alternatively, these species can be retained, in which case losses will be incurred during the process of reconciliation.

Which strategy is more appropriate for a particular analysis depends on how the data were sampled. If the sequences of interest were derived from species with good quality, whole genome sequence, the failure to observe a gene family member in one of these species is suggestive of a true gene loss. In this case, species that are not associated with any gene should be retained, so that losses are incurred during reconciliation. However, if the species represented in the gene tree do not have reliable whole genome sequences, then pruning the species tree prior to reconciliation may be appropriate.

With this in mind, Notung offers two approaches to pruning the species tree. By default, Notung finds the lowest common ancestor (LCA) of the species that are associated with genes in the gene tree, and only prunes those species that diverged prior to the LCA. All species within the clade rooted at the LCA are retained. This strategy is designed for high quality data sets, where the absence of a gene is indicative of gene loss. Alternatively, the user may opt to prune all species that are not associated with at least one gene. In the GUI, this is achieved by selecting the "prune" check-box in the Reconciliation panel. Please see Chapter 12.2 - Prune on page 104 for command-line options regarding species tree pruning. This strategy is appropriate when the failure to observe a gene in a particular species may be due to sparse data or experimental error. This approach is also appropriate in cases where

the species tree was constructed for some other purpose and contains species that are not of interest in the current analysis.

Unlike species nodes, which may be associated with zero, one, or more than one gene, each gene must be associated with exactly one species. If Notung cannot identify the species associated with one or more leaves in the gene tree, it will print a diagnostic message and the reconciliation process will be terminated. This could indicate a species tag formatting problem. However, in some cases, the leaf set in the gene tree is a superset of the gene family of interest in the current study. For example, the user may have started with a broad analysis, but now wishes to restrict the study to a smaller set of species. In this case, it is convenient to be able to prune from the gene tree, those leaves that are not associated with a species in the restricted set. Notung's command line interface offers a utility for this purpose that can be used to obtain a pruned gene tree in a preprocessing step. See Section 12.2 - Prune on page 104 for more information.

5.2 Reconciling Binary Trees with Transfers

In Notung, the user can reconcile gene trees with species trees to infer transfer events (*e.g.* horizontal/lateral gene transfer), along with duplications, losses, and co-divergences. We call the model with transfer events the DTL-model, or DTLI-model if the species tree is non-binary. Currently, analyses performed on non-binary gene trees with the DTL model are restricted to the command-line interface. See Section 12.8 - Non-binary gene trees with the DTL model on page 117 for more information. Analyses with binary gene trees can be performed either in the GUI or on the command-line. To enable the DTL-model or DTLI-model, the user needs to select the "Infer Transfers" option in the Reconciliation panel. By doing that, the user can specify a cost value for a transfer event. The event costs can be set by clicking the "Edit Values" button on the bottom right corner. The default cost values are 1.5 for Duplication, 3 for Transfer, and 1 for Loss. Once "Infer Transfers" is selected, operations like reconciliation and rooting will be performed to minimize the events cost using the DTL-model, in other words, with possible transfer events taken into consideration. For a more detailed description of the DTL-model, please refer to Chapter 4.

After performing a reconciliation with the DTL-model, Notung will display a binary gene tree with events labeled on it. Inferred Duplication and Co-Divergence events will be indicated by red 'D' and 'cD' above corresponding internal nodes respectively, and gray external nodes (leaf nodes) with '_LOSS' in their name labels indicate Loss events. For an inferred transfer event, the gene tree edge associated with that transfer (transfer edge) will be labeled in yellow, and a yellow triangle, with the letter 'T', will be displayed on the middle of that transfer edge (see Figure 5.3(a)). The yellow triangle on each transfer edge is clickable. When clicked, the species donor and species recipient of that transfer will be displayed on top of that

transfer edge, in the format of 'from donor species to recipient species.' Alternatively, the user can also view the species information by selecting "**Display Internal Node Species Names**" and "**Display Leaf Node Species Names**" under "**Display Options**" on the menu bar. Information about all inferred transfers, along with other statistics related to this reconciliation can be viewed in a pop-up window by selecting "**Event Summary**" or "**Parsable Statistics**"from the "**About This Tree**" menu.

5.2.1 Temporal Feasibility

In reconciliation with the DTL-model, it is important to beware of the temporal feasibility of a solution, as each inferred transfer will impose temporal constraints. For example, a transfer implies both its donor and recipient species co-existed in a certain temporal interval. When there are multiple transfer events inferred, it is possible that the multiple transfers impose conflicting temporal constraints. We call such problematic reconciliation solutions (temporally) infeasible solutions. Thus, it is crucial to check the temporal feasibility when transfer is included in the reconciliation model. For theoretical background about temporal feasibility, please see Chapter 4.1.1 - Temporal Feasibility on page 28.

By default, Notung checks the temporal feasibility whenever an event history is performed with the DTL-model, to ensure that any solution that is presented to the user is temporally feasible. If all the solutions with minimal Event Cost are temporally infeasible, a dialog box will pop up (GUI mode) or a warning message will appear on the screen (command line mode), warning the user of the issue. In that case, no reconciliation solution will be shown in the GUI, and no reconciliation will be output in the command line mode. Please be aware of the fact that all minimum-cost solutions are infeasible does not mean that a temporally feasible solution with that particular set of event costs does not exist. A sub-minimum-cost solution may be temporally feasible.

5.2.2 Multiple Solutions

When transfer is included in the reconciliation model, it is possible for a reconciliation to have more than one minimum-cost solution with a given set of event costs. For theoretical background about multiple solutions, please see Chapter 4.1.1 - Multiple Solutions on page 27. In the GUI, after a reconciliation is performed with the DTL or DTLI-model, the number of feasible solutions will be displayed in the status bar on the bottom of the GUI, along with the inferred number of events and total cost. The number of feasible solutions will also appear in the pop-up window if the user selects "**Event Summary**" or "**Parsable Statistics**" from the "**About This Tree**" menu. In the GUI, a green circle on a node in the gene tree panel indicates the existence of multiple solutions within the subtree rooted at that particular node.

It is easy to browse through all feasible solutions by clicking the corresponding green circle(s) repeatedly (see Figure 5.2).

5.2.3 Rearrange and Resolve with Transfers

Currently rearranging and resolving with transfers (DTL-model) are not supported in the Notung GUI. If the optional **Infer Transfers** from the Reconciliation panel is selected, clicking **Perform Rearrangement** from the **Rearrange** panel, or clicking **Resolve Polytomies** from the **Resolve** panel, will bring up a pop-up window with warning messages, reminding the user that rearranging with transfer is only supported via the command-line.

5.3 Reconciling Non-Binary Trees

Notung can reconcile binary gene trees with non-binary species trees, as well as non-binary gene trees with binary species trees. The differences between these functions and traditional reconciliation of binary gene trees with binary species trees are summarized briefly here. For a more detailed discussion of reconciliation with non-binary trees, see Chapter 4.2 - Non-Binary Trees on page 29. Note that orthologs, paralogs, and xenologs can only be inferred on binary gene trees reconciled with binary species trees.

Reconciling a binary gene tree with a non-binary species tree results in a binary gene tree with duplications, transfers, and losses added. Notung distinguishes between cases in which disagreement can only be explained by a gene duplication or transfers and cases in which the disagreement may be due to co-divergence. When reconciling a gene tree with a non-binary species tree, duplications appear in the tree as small red squares with red D's, and transfers appear as yellow edges with yellow triangles and T's, while co-divergences, other than binary speciation events, are small pink squares with pink cD's (see Figure 5.2 and Figure 5.3).

If two or more orthologous genes are missing from species that are children of the same polytomy, then it is more parsimonious to infer a loss of the common ancestor of those genes. We refer to such losses as **polytomy losses**. For example, in Figure 5.1, members of the hypothetical Y gene family are missing from two species, bandicoot and opossum. These species are children of the same polytomy in the species tree in Figure 4.1. Notung infers a single loss, labeled with the names of species from which the gene is absent, as well as the label of the corresponding polytomy in the species tree. By default, polytomy losses are labeled with the species that lack the gene. However, if a polytomy loss is associated with many sibling species, the default display can produce very long labels. Users can instead opt

to label polytomy losses with the number of species in which the loss occurred, as well as the label and the total number of children of the polytomy, illustrated in Figure 5.3(b).

Reconciling a non-binary gene tree with a binary species tree results in a non-binary, reconciled gene tree. A reconciled, binary gene tree can be obtained by using the Resolve function (see Chapter 8 - Resolve Mode on page 81).

> *NOTE*: In this case, transfers can currently be inferred only via the command-line interface, as discussed in Section 12.8.

Reconciliation of a non-binary gene tree with a binary species tree differs from binary reconciliation in two important ways. First, a polytomy in a non-binary gene tree may be annotated with more than one event. For example, the reconciled non-binary gene tree in Figure 5.4(a) has a polytomy annotated with two duplications and a loss. Recall that a gene tree polytomy is an indication that although its children evolved by successive binary divergences, the order in which the taxa diverged is unknown. Since this binary branching pattern is unknown, the relative order of events with respect to those divergences cannot not be determined, either. The polytomy in Figure 5.4(a) communicates that at least two duplications and one loss occurred in the subtree descending from the polytomy, but the exact timing of those events is unknown. See Chapter 4.2 - Non-Binary Trees on page 29 for a detailed explanation of duplications and losses in reconciled non-binary gene trees.

Second, there may be several alternate hypotheses for the reconciliation of a non-binary gene tree. Since the true binary branching pattern of a polytomy is unknown, Notung infers events for all binary resolutions with the minimal Event Score. If there is more than one optimal binary resolution, multiple reconciliations will result. Notung addresses this issue by presenting all alternate event histories to the user. Each event history represents a different combination of duplications and losses that could result in the same minimal DL Score. Initially, Notung arbitrarily selects one event history to present in the tree panel. The other optimal histories may be viewed using the drop-down menu labeled "**Select an optimal event history**," as shown in Figure 5.4. This menu gives a list of up to 50 optimal event histories. If there are more than 50 optimal event histories, they can be generated using the Command Line Interface (see Chapter 12 - Command Line Options and Batch Processing on page 98). For a more detailed discussion of alternate event histories, see Chapter 7 - Rearrange Mode on page 73.

5.4 Using Reconciliation Commands

To regulate reconciliation options and behaviors:

A number of different options affecting the reconciliation algorithm and results are now available, including the ability to infer transfers.

- Click the appropriate checkbox in the Reconciliation task panel.

  √ **Infer Transfers (default: OFF)** When this box is checked, transfers are included in the optimization criterion. To infer transfers with duplications, check this box.

  NOTE: Duplications are always included in the optimization. To consider only transfer events as a source of tree disagreement, you may simply set the cost of duplication to a very high value relative to the cost of transfers.

  √ **Prune (default: OFF)** When this box is checked, the species tree is "pruned" to remove taxa that are not represented in the gene tree.

  NOTE: By default, Notung does not prune the species tree. This assumes that genes missing in species are due to true loss events and not incomplete taxon sampling. As a result, spurious losses may be inferred if the species tree contains taxa that were insufficiently sampled for the genes under consideration.

To change the parameter values:

1. Click the "**Edit Values**" button. A dialog box appears.
2. Enter the appropriate values in the text field, and then click "**Apply Changes.**"

NOTE: This will change the value settings only for the tree that is currently selected. Each history state saves the parameter values used at that state; when moving through the history, parameter values may change depending on the state and tree viewed. For more information on history states, see Chapter 9 - History on page 85.

NOTE: Transfer costs can only be edited if the "Infer Transfers" option is selected.

To reconcile a gene tree with a species tree:

1. Click the **Reconciliation** tab to enter Reconciliation mode.
2. Click the "**Reconcile/Rereconcile**" button. A dialog box appears.

3. In the dialog box, select the correct species tree in the drop-down menu.

4. Check that Notung correctly identified the species naming convention used in the gene tree. The available settings are:

   - Prefix of the gene label (*i.e.*, SPECIESGENE)
   - Postfix (underscore required) of the gene label (*i.e.*, GENE_SPECIES)
   - NHX: species label is stored in the NHX comment field in the gene tree file. (*i.e.*, GENENAME[&&NHX:S=SPECIES])

   If the convention selected by Notung is not the naming convention used in the gene tree, change it by selecting the appropriate radio button. See Appendix A.4 - Specifying the Species Associated with Each Gene on page 129 for details about species tag specifications.

   NOTE: The Prefix and Postfix formats require species names to be embedded in the gene names. NHX Species Tag format embeds the species information in a Newick comment field. When this format is used, the information will not appear on the screen unless the "**Display Leaf Node Species Names**" option in the **Display Options** menu is selected (see Chapter 11.1 - Display Options on page 93).

5. In the dialog box, click "**Reconcile**."

The reconciled tree appears in the Tree Panel (see Figure 5.1(b)). Internal nodes of the tree represent events that caused a bifurcation. In a reconciled gene tree, the inferred events appear as:

Speciation Node is a small blue square with no text.

Duplication Node is a large red square, displayed with a red 'D'. In a non-binary gene tree, the number of duplications associated with a polytomy will also be shown with a red D (*e.g.*, Figure 5.4(a)).

Co-divergence Node is a pink square with the letters "cD"; only inferred and shown if the species tree is non-binary.

Transfer The edge is highlighted in yellow. A yellow triangle, with a yellow 'T', appears halfway down the edge. Transfer events are represented on tree branches to indicate the bifurcation that resulted from the transfer, as well as the recipient of the transfer. When the triangle is clicked, the donor and recipient taxa of that event are displayed.

Loss The leaf node appears in light gray type is labeled "<species>*LOST", indicating the species in which the loss occurred.

A status bar at the bottom of the program window (Figure 5.3) provides additional information about the reconciliation. This includes the total cost of the reconciliation (Event Score), as well as a summary of the number of inferred events, the total number of optimal, feasible solutions, and the species tree used for the reconciliation. In addition to displaying events visually, Notung-2.9can also generate a detailed report on event histories in a text format. See Section 5.5 - Additional Reports on page 51 for more information.

The species associated with the ancestral nodes in the reconciled gene tree can also be graphically displayed on the tree. The user can view this information by selecting "**Display Internal Node Species Names**" and "**Display Leaf Node Species Names**" under "**Display Options**" on the menu bar.

To hide losses/duplications/transfers/co-divergences:

The duplication, co-divergence, or transfer marks or loss nodes can be hidden to avoid a cluttered image.

- Click the appropriate checkbox in the Reconciliation task panel.

  √ **Display Loss Nodes (default: ON)** When this box is checked, the inferred lost nodes appear in the tree. To hide inferred loss nodes, uncheck the box.

  NOTE: When you uncheck "**Display Loss Nodes,**" Notung will reset the image so that the whole tree fits in the tree panel.

  √ **Display Duplications (default: ON)** When this box is checked, duplications are indicated on internal nodes by red squares and D's. To make the red D's disappear, uncheck the box.

  √ **Display Co-Divergences (default: ON)** When this box is checked, co-divergences are indicated on internal nodes by pink squares and cD's. To make the pink cD's disappear, uncheck the box.

  √ **Display Transfers (default: ON)** When this box is checked, transfers are indicated on highlighted edges by yellow triangles and T's. To make the yellow T's disappear, uncheck the box.

Options that are not currently available are displayed in gray type to indicate that they are disabled. In particular, the above options will be grayed out if no reconciliation has been performed. The "**Display Co-Divergences**" option will also be displayed in gray if the gene tree was reconciled with a binary species tree, and the "**Display Transfers**" option will be grayed out if the reconciliation did not consider transfers.

To display the number of species in polytomy losses:

By default, polytomy losses are labeled with the names of the species from which they are absent.

1. Go to the "**Display Options menu**."
2. Click the "**Use Species Names in Polytomy Losses**" box.

   This causes polytomy losses to be labeled with the number of children of the polytomy lost, the total number of children of the polytomy, and the name of the polytomy in which these losses occurred.

To display the donor and recipient of transfers:

The donor and recipient species of transfers in the gene tree can be displayed by clicking on any transfer triangle in the gene tree. Clicking a triangle again will hide this information. If the *gene tree is non-binary*, or if *transfers* were inferred, there may be more than one optimal reconciliation.

To view alternate optimal event histories for non-binary gene trees:

If more than one optimal event history exists for a reconciled, non-binary gene tree, the drop down menu, "**Select an optimal event history**," will be enabled.

- From the drop-down menu, select an alternate event history.

  The tree panel will now show a new tree corresponding to the selected alternate history.

If there is only one optimal history or if the tree has not been reconciled, the drop down menu will be grayed out.

NOTE: When a user selects a different alternate event history from the "Select an optimal event history" list, Notung rebuilds the tree from data saved at the time of reconciliation. Any manual swaps made to a previously viewed event history will be lost. Therefore, if you wish to save information after a manual swap, you must save your tree. See Chapter 3.3 - Opening and Saving Trees on page 15 for more information.

To view alternate optimal event histories when inferring transfers:

If more than one optimal event history exists for a tree reconciled with transfers, green circles will appear around nodes in the gene tree.

- On the gene tree, click on the nodes circled in green. The tree will now show an alternative optimal solution.

Clicking on a node will loop through the different alternate histories affecting areas below the node. To loop through all optimal histories, click on the green circle(s) closest to the root of the tree.

> *NOTE*: When a user moves to a different mode, Notung rebuilds the tree from data saved at the time of reconciliation. Any manual click to change the solution will be lost. Therefore, if you wish to save information after a manual change, you must save your tree. See Chapter 3.3 - Opening and Saving Trees on page 15 for more information.

To undo the reconciliation:

- Click the "**Unreconcile**" button. The most recent reconciliation will be undone. This option is grayed out if the gene tree has not been reconciled.

To display ancestral associations:

1. Click "**Display Options**" from the menu bar.
2. Select "**Display Internal Node Species Names**"

To display the species tree:

1. Click the "**Show pruned species tree**" button. A dialog box appears.
2. Enter a title in the text field and click "**OK**."

This option is grayed out if the gene tree has not been reconciled.

5.5 Additional Reports

5.5.1 Event Summary

Notung can provide detailed information on the inferred event histories in text format. These reports include information on the timing of duplications (upper and lower bounds on the time of duplication), transfers (the species from which the gene originated and the which received the gene), and losses (the species in which the loss occurred), as well as timing of co-divergences with the gene tree. Transfers, insertions, duplications, co-divergences, and bounds are identified by internal node names in these reports. For losses, each node in the species tree is listed, followed by the number of losses associated with that species. Finally, a table of species is included, where rows represent each node in the species tree, and values in a row indicate the number of duplications, transfers from, transfers to, and losses associated with that species. The files, and the information provided, are described in detail below.

> When viewing this information, it may be helpful to display the names of internal nodes in the gene tree, as well as the species associated with these nodes. To do so, from the "**Display Options**" menu (see Chapter 11.1 - Display Options on page 93), turn on "**Display Internal Node Names**" and and "**Display Internal Node Species Names**".

The Event Score of the reconciled tree appears at the top of the window.

This is followed by a table on duplication bounds, described in three columns. The internal node representing the duplication is listed in the first column. Duplication bounds are described in the next two columns: lower and upper bounds, respectively, relative to the species tree, expressed as node names in the species tree. The total number of duplications appears below this table.

If the *species tree is non-binary*, co-divergence bounds, if any, are described in the three columns below duplication bounds. The left column gives the internal node in the gene tree where the co-divergence occurred. The next two columns provide the lower and upper bounds, respectively, on the species tree node in which the event (speciation or allelic divergence) may have occurred. The total number of co-divergences is listed below this table.

If included in the set of events, information on transfers will appear in a table below these bounds. The first two columns list the nodes in the gene tree where the transfer occurred, representing where the transfer originated and ended, respectively. The timing of the transfer event is listed in the two subsequent columns, respectively: the species from which the transfer

came (donor), and the species which accepted the transfer (recipient). The total number of insertions is listed below this table.

Information on losses follows. Rather than reporting each loss separately, losses are summarized according to the taxa in which they occurred. This table has two columns. The first column is a list of all the nodes in the species tree; the second column gives the number of inferred gene losses that occurred in that species. Polytomy losses are assigned to the corresponding polytomy, rather than the individual species which lack the gene. For example, the polytomy loss in Figure 5.3 is reported as a single loss in *Metatheria*. The total number of losses is listed below this table.

Following information on each separate event type, there is a table providing total event counts per species. The first column lists all the nodes in the species tree. Subsequent columns give the number of times each species is associated with an events: duplication, co-divergence, transfer donor, transfer recipient, and loss. The total number of each event is listed, again, below this table.

To show information on events:

- Select the "**Event Summary**" menu item from the "**About This Tree**" menu. A new window will appear.

This option is grayed out if the gene tree has not been reconciled.

Parsable Statistics

This file is designed to contain much the same information as the "**Event Summary**" file, but in a format that is easily parsed by a computer. The first line is a tab-delimited set of values, with the header on the line below.

This line includes: the number of duplications (nD), co-divergences (nCD), transfers (nT), and losses (nL); the size of the gene tree without losses in terms of the number of leaves (|L(G)|) and total nodes (|G|), as well as the size of the associated species tree in terms of total nodes (|S|); the heights of the gene (h(G)) and species (h(S)) trees in terms of the longest path from root to leaves; whether the root of the gene tree is mapped to the root of the species tree (isRoot), the species to which the gene tree root is mapped (root(G)), and the root of the species tree (root(S)); the range of edge weights (minEW, maxEW); the number of optimal roots (Roots) and reconciliations, both candidate (Cand) and feasible (Feas); the cost of duplication (cD), co-divergence (cCD), transfer (cT), and loss (cL); the size of the largest polytomy (largest Poly) and the number of polytomies (numPoly).

The remainder of the file provides the same information as "**Event Summary**." The lines are tab-delimited and start with `#<x>` to distinguish the type of information displayed. `<x>` can be any of D, CD, T, L, S, or R. #D demarcates information on duplications, #CD demarcates information on co-divergences if the species tree is non-binary, #T demarcates information on transfer, and #L demarcates information on losses tallied per species. Lines in the final summary table are demarcated with #S. The final information in this file begins with #R. It provides the species edge, defined by the species node below the edge, on which the gene tree is rooted.

To view tree statistics and event details in an easily parsable format:

- Select the "**Parsable Statistics**" menu item from the "**About This Tree**" menu. A new window will appear.

This option is grayed out if the gene tree has not been reconciled.

5.6 Inferring Orthologs, Paralogs, and Xenologs

Many types of comparative analysis require the identification of different types of homologs. Gene tree - species tree reconciliation is the most reliable and robust approach to distinguishing orthologs from paralogs. Reconciliation is also increasingly used to infer horizontal gene transfer events, and thus, can be used to identify xenologs.

Notung utilizes information from a reconciled gene tree to infer the homology relationships between all pairs of leaves in the gene tree. Currently, homology inference in Notung is only supported for the case where both the gene and species trees are binary. An extension to non-binary trees will be included in a future version.

Notung's homology classification functions can be accessed from the GUI or the command line. The Notung GUI contains an interactive feature in the **Reconciliation** task panel that allows the user to investigate homology through a point and click interface (Section 5.6.2). In addition, Notung can output a table summarizing the homology relationships between all pairs of leaves in a reconciled gene tree (Section 5.6.3). This homology table can be saved from either the GUI or the command-line interface. The batch processing feature of the command-line interface also allows for automated generation of homology tables for a large number of gene trees. See Chapter 12 - Command Line Options and Batch Processing for more information.

Homology Classification in Notung. Notung uses the *cenancestor-based* homology terminology introduced by Fitch [12]:

> *Orthologs*: "two homologous characters whose common ancestor lies in the cenancestor[1] of the taxa from which the two sequences were obtained."
>
> *Paralogs*: "two homologous characters arising from a duplication of the gene for that character."
>
> *Xenologs*: "two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters."

These definitions differ from the more commonly used *event-based* definitions, where a pair of genes are:

> *Orthologs*: two homologous characters that diverged via a speciation event.
>
> *Paralogs*: two homologous characters that diverged by a duplication event.

However, under the duplication-loss model (**Infer Transfers OFF**), the cenancestor-based definitions and the event-based definitions are identical.

In Section 5.6.1 - Homology Terminology, we discuss the definitions that Notung uses for homology classification, with particular attention to homology classification under the duplication-transfer-loss model (**Infer Transfers ON**). If your goal is to distinguish between orthologs and paralogs in gene families in which horizontal gene transfer does not occur, you may wish to skip this section and go on to Section 5.6.2.

5.6.1 Homology Terminology

Notung employs the cenancestor-based definitions, stated above, for homology classification because they present no ambiguity when distinguishing xenologs from orthologs. In a gene family in which gene duplication is the only mechanism by which new copies of a gene family arise, the commonly-used event-based definitions are unambiguous: Every pair of genes in a reconciled tree can be assigned to exactly one of these two classes: orthologs or paralogs.

[1] The *cenancestor* of a pair of genes is the lowest common ancestor (LCA) of the species containing that pair.

However, in a gene family with horizontal gene transfer, the event-based definitions of ortholog and paralog can overlap with the definition of xenolog. Consider, for example, a gene family in which a speciation event gave rise to two gene lineages, one of which subsequently experienced a horizontal transfer. Since their shared history includes a horizontal transfer, they are xenologs. However, according to the event-based definition of orthologs, the resulting gene pair are also orthologs. For example, in Figure 5.5, the speciation event at node 2 gave rise to genes $\widehat{g}_X$ and g_Z. These genes are xenologs because their history, since their LCA, involves a horizontal transfer. However, according to the event-based definition, they are orthologs because the event that caused the divergence at node 2 was a speciation.

The cenancestor-based definitions remove this ambiguity. Under that definition, genes $\widehat{g}_X$ and g_Z are *not* orthologs, because their LCA (node 2) is not in their cenancestor (species β). Note that the cenancestor-based and event-based definitions are equivalent in a pair of genes that evolved by duplication and loss only (either no transfers inferred or **Infer Transfers OFF**). In the absence of transfers, the LCA of two genes lies in their cenancestor if and only if there was a speciation event at the LCA.

As this example illustrates, Fitch's cenancestor-based definitions result in a precise and non-overlapping classification of xenologs and orthologs. However, a consequence of these definitions is that xenologs encompass a broad set of gene pairs with highly diverse properties and relationships.

First, a pair of xenologs may have arisen from any event — transfer, speciation, or duplication — at the LCA of the gene pair. For example, in Figure 5.5, the three gene pairs $(\widehat{g}_X, g_Y)$, $(\widehat{g}_X, g'_Y)$, and $(\widehat{g}_X, g_Z)$ are each xenologous. The event at the LCA of genes $\widehat{g}_X$ and g_Y is a transfer (node 1). The event at the LCA of genes $\widehat{g}_X$ and g_Z, is a speciation (node 2). The event at the LCA of genes $\widehat{g}_X$ and g'_Y is a duplication (node 4).

Second, xenologous pairs, even in the same species, may vary greatly in how closely they are related; the species associated with a pair's LCA may post-date, pre-date, or be the same as their cenancestor. Figure 5.5 illustrates these three possibilities. Genes $\widehat{g}_X$ and g_Y are xenologs whose LCA (node 1) post-dates the cenancestor (species β). For xenologs $\widehat{g}_X$ and g_W, their LCA (node 6) pre-dates the cenancestor (species γ). And in a third example, for xenologs $\widehat{g}_X$ and g_V, their LCA (node 7) is in their cenancestor (species α).

Xenolog Classification. In order to achieve a more informative classification within the broad class of xenologs, we have developed xenolog subtypes that capture distinctions in how horizontally transferred genes are related [4]. Under the duplication-transfer-loss (DTL) model (**Infer Transfers ON**), Notung infers orthologs, paralogs, and xenologs, and further labels xenologs according to this subclassification. Here, we summarize these xenolog subclasses for a gene family in which a single horizontal transfer has occurred. The inference of

subclasses for a gene family history with an arbitrary number of transfers and duplications is described in detail in Darby *et al.* [4]. The theoretical properties of the classification, pseudocode, and examples of its application to real data are also described in that publication.

Our classification groups xenologous gene pairs into five disjoint subclasses (Table 5.1). Recall that a pair of genes are xenologous, if they are related through a transfer event at some point in their shared history. In other words, the path between the two genes will pass through the transfer. Let $\widehat{g}$ and g be genes such that there is a transfer on their path and gene $\widehat{g}$ is a descendant of the transfer recipient. We say that gene g is a xenolog of $\widehat{g}$ and treat $\widehat{g}$ as the reference.

		P	Paralogs
		O	Orthologs
Xenologs	Class	PX	Primary xenologs
		SDX	Sibling donor xenologs
		SRX	Sibling recipient xenologs
		OX	Outgroup xenologs
		IX	Incomparable xenologs
	Suffix	-X′	-autoxenolog
		-Xp	-paraxenolog

Table 5.1: Abbreviations for different homology classes.

Gene g is a *primary xenolog* (PX) of $\widehat{g}$ if the divergence at their LCA in the gene tree was caused by a transfer event. For example, g_Y is a primary xenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g_Y)$ is designated PX) because their LCA (node 1) was the transfer donor. Otherwise, we classify the xenologs based on the species containing g. If g is in a species that is more closely related to the donor species than the recipient species, then g is a *sibling donor xenolog* (SDX) of $\widehat{g}$. Similarly, if g is in a species that is more closely related to the recipient species than the donor species, then g is a *sibling recipient xenolog* (SRX) of $\widehat{g}$. Otherwise, g is in a species that diverged before the transfer cenancestor (the LCA of the donor and recipient species) and is equally related to the donor and recipient species; in this case, g is an *outgroup xenolog* (OX) of $\widehat{g}$. In Figure 5.5, g_Z is a sibling donor xenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g_Z)$ is SDX); g_W is a sibling recipient xenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g_W)$ is SRX); and g_V is an outgroup xenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g_V)$ is OX). Sometimes, g will be in the same species as $\widehat{g}_X$. We add the extra annotation of *autoxenolog* (X′) to note these special cases. For example, g_X is a sibling recipient autoxenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g_X)$ is SRX′).

While Fitch's cenancestor-based definitions distinguish between orthologs and xenologs, it is still possible for a pair of genes to be both xenologs and paralogs, i.e., if g is a xenolog of $\widehat{g}$ and the genes diverged at a duplication node. We introduce the term *paraxenolog* (X^P) to account for such cases. A paraxenolog can also be a member of one of the standard classes (above) and will be assigned that class. For example, in Figure 5.5, gene g'_Y is a sibling donor

paraxenolog of $\widehat{g}_X$ (pair $(\widehat{g}_X, g'_Y)$ is SDX^P) since there was a duplication at their LCA (node 4).

Note that because of the directional nature of transfers, the classification of xenolog pairs is not symmetric, and we say that g is a xenolog of $\widehat{g}$. In a table with genes for columns and rows, the entry for pair $(\widehat{g}, g)$ will contain the annotation, and the entry for pair $(g, \widehat{g})$ will be $*$.

When there are multiple transfers on the path between genes $\widehat{g}$ and g, it is possible that $\widehat{g}$ will be descended from one transfer and g will be descended from another transfer. In this case, g is a xenolog of $\widehat{g}$, and $\widehat{g}$ is a xenolog of g. We say that $\widehat{g}$ and g are *incomparable xenologs* (IX). Both entries in the table — $(\widehat{g}, g)$ and $(g, \widehat{g})$ — will contain the annotation. Xenolog classification in the presence of multiple transfers is discussed in detail in [4].

5.6.2 Interactive Homology Mode

From the Notung GUI, users can investigate the homologs of any contemporary gene through a point and click interface. The information displayed can then be saved as an image in PNG format. We demonstrate how to get a table containing this information in Section 5.6.3 - Homology Table.

Here, we demonstrate how Notung's point and click interface can be used to investigate homology relationships using two examples. One example shows an analysis with duplications and losses. The second analysis demonstrates homology classification in a gene tree with both horizontal transfers and duplications. Trees for both examples are included in the Notung distribution and can be found in the `sampleTrees` directory.

Example 1: Investigating Orthologs and Paralogs. Functions for inferring homology relationships when using the DL event model (**Infer Transfers OFF**) are illustrated here.

1. Select the Reconciliation task panel and reconcile the gene tree, if it has not yet been reconciled.

2. Click on the "**Show Homology**" button to enter the interactive Homology mode. The homology mode legend appears in the upper-left corner of the tree panel.

3. Click on or mouse over a leaf in the gene tree. It will be highlighted in light blue. Orthologs of this gene will be highlighted in dark blue and paralogs in pink, as shown in Figure 5.6.

4. To save an image displaying these relationships, select "**File → Save Current View as Image (PNG)**" from the menu.

   NOTE: The image will contain the Homology legend, as well as the orthologs and paralogs of any currently selected gene. Currently, "**File → Save Whole Tree as Image (PNG)**" does not show homology relationships.

5. To minimize the legend, click on "**hide**" in the legend. Click on the minimized legend to show the full legend again.

6. The legend can be dismissed entirely by clicking "**close**." If you re-enter Homology mode, the legend will be visible again.

7. To exit Homology mode, click the "**Show Homology**" button.

Example 2: Investigating Orthologs, Paralogs, and Xenologs. Functions for inferring homology relationships when using the DTL event model (**Infer Transfers ON**) are illustrated here. This example reflects the scenario presented for xenolog classification in Figure 5.5.

1. Select the Reconciliation task panel and reconcile the gene tree with **Infer Transfers ON**, if it has not yet been reconciled.

2. Click on the "**Show Homology**" button to enter the interactive Homology mode. The homology mode legend appears in the upper-left corner of the tree panel.

3. Click on or mouse over a leaf in the gene tree. It will be highlighted in light blue. Orthologs of this gene will be highlighted in dark blue, paralogs in pink, and xenologs in yellow, as shown in Figure 5.7.

   NOTE: Specific colors are not associated with the different xenolog subclasses; however, this information is available in the homology table, as shown in Table 5.3. See Section 5.6.3 - Homology Table on page 59 for instructions on obtaining this table.

4. To save an image displaying these relationships, select "**File → Save Current View as Image (PNG)**" from the menu.

   NOTE: The image will contain the Homology legend, as well as the orthologs, paralogs, and xenologs of any currently selected gene. Currently, "**File → Save Whole Tree as Image (PNG)**" does not show homology relationships.

5. To minimize the legend, click on "**hide**" in the legend. Click on the minimized legend to show the full legend again.

6. The legend can be dismissed entirely by clicking "**close**." If you re-enter Homology mode, the legend will be visible again.

7. To exit Homology mode, click off the "**Show Homology**" button.

5.6.3 Homology Table

Notung can also provide information on homology relationships between all pairs of leaves in a gene tree in tabular format. This *homology table* is a matrix with a row (resp. column) for every leaf in the gene tree. The $[g_i, g_j]$ entry in the table gives the homology relationship of genes g_i and g_j. Orthologous and paralogous genes are indicated by an "O" and a "P" in the table, respectively. Entries for xenologous pairs contain an "X". Notung further classifies xenologs into subtypes; information about each subtype, as described in Section 5.6.1, is provided in the table as well. Abbreviations for homolog types are shown in Table 5.1.

> *NOTE:* Xenolog pairs are not symmetric. Ordered pairs that are labeled with a star will be classified in the other direction. For example, in Table 5.3, entry `(gHat_X,g_Y)` is `PX`, meaning `g_Y` is a primary xenolog of `gHat_X`. Entry `(g_Y,gHat_X)` is `*`. For more information, refer to Section 5.6.1.

Notung can output the homology table in three table formats: tab-separated, comma-separated (CSV), or HTML formats. HTML tables are coded to color cells according to their assigned homology relationship. Cells representing orthologs have a blue background, cells representing paralogs have a pink background, and cells representing xenologs have a green background.

> *NOTE:* Tab-delimited tables can usually be pasted directly into spreadsheet applications like Excel. CSV formatted tables can be opened by most spreadsheet programs via the file menu. HTML format tables can be pasted directly into web pages.

To view the Homology table:

1. Select the Reconciliation task panel and reconcile the gene tree, if it has not yet been reconciled.

2. Go to the "**About This Tree**" menu.

3. Select the "**Homology Table**" option with the desired format (**CSV**, **Tab delimited**, or **HTML**) in the pull-down menu. The selected table will be displayed in a pop-up dialog box, for example as in Table 5.2 and Table 5.3.

   NOTE: To save a copy of the homology table, in the pop-up box click "**Copy to clipboard**," then paste to a text buffer of your choice and save.

The homology table for the tree in Figure 5.6, reconciled under the DL event model (**Infer Transfers OFF**), is illustrated in Table 5.2. The homology table for the tree in Figure 5.7, reconciled under the DTL event model (**Infer Transfers ON**), is illustrated in Table 5.3. This example reflects the scenario presented for xenolog classification in Figure 5.5.

`Homology Table for: genetree_SMALL`

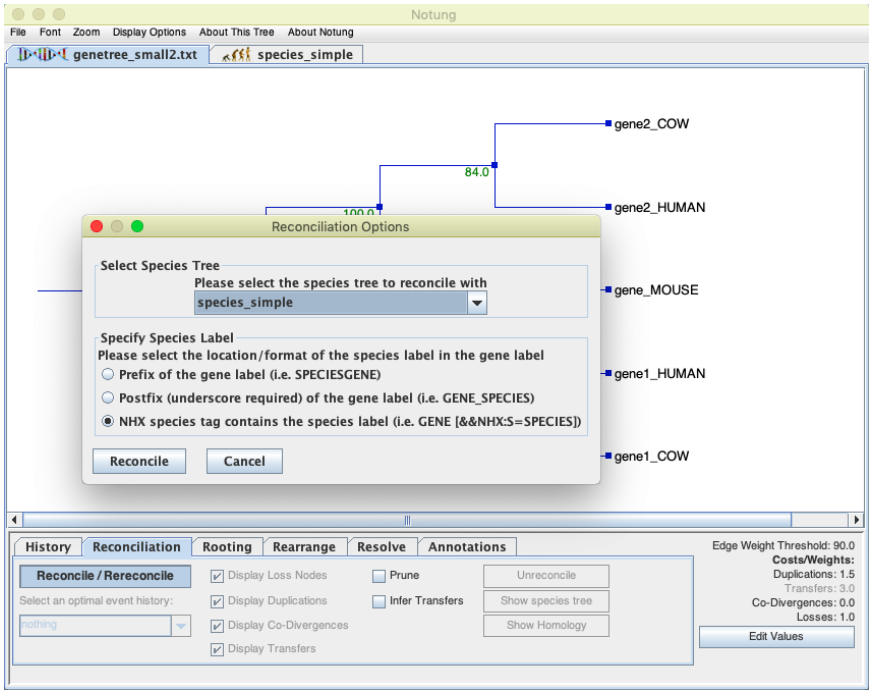
	gB_human	gA_human	gA_mouse	g_gorilla	gB_mouse	gY_cow	gX_cow
gB_human	.	P	P	P	P	O	O
gA_human	P	.	P	P	P	O	O
gA_mouse	P	P	.	O	P	O	O
g_gorilla	P	P	O	.	P	O	O
gB_mouse	P	P	P	P	.	O	O
gY_cow	O	O	O	O	O	.	P
gX_cow	O	O	O	O	O	P	.

Table 5.2: Homology table, showing orthologs and paralogs, for the example from Figure 5.6. In this example, `genetree_SMALL` has been reconciled with `speciestree_SMALL` using the DL event model (**Infer Transfers OFF**). The abbreviation key is provided in Table 5.1. Cells at the intersection of the column and row representing the same gene are labeled with a dot.

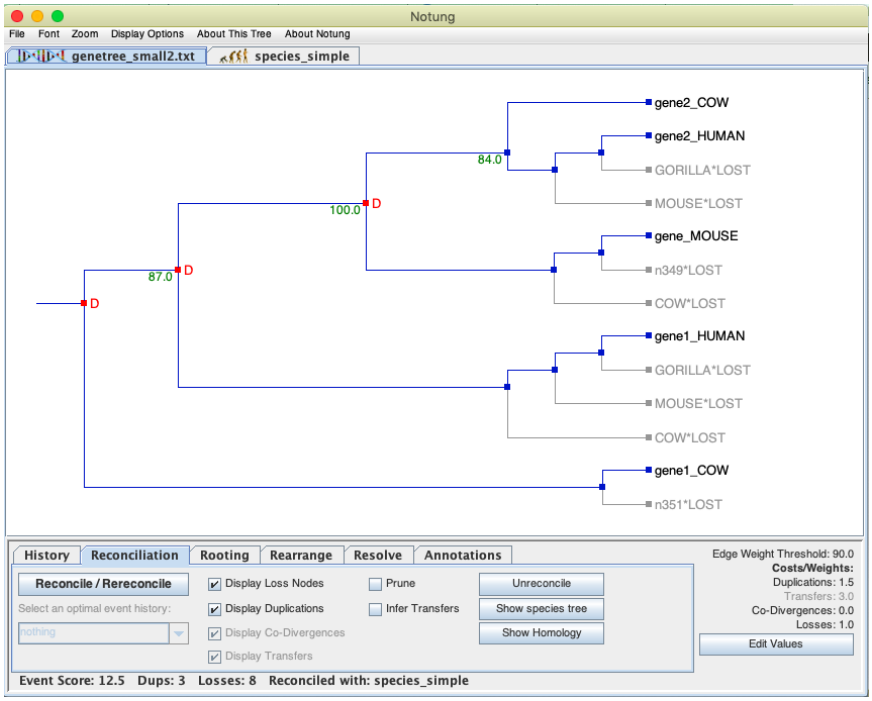
Homology Table for: xenolog_gene

	g_Y	gHat_X	g_Z	gPrime_Z	gPrime_Y	g_X	g_W	g_V
g_Y	.	*	O	P	P	O	O	O
gHat_X	PX	.	SDX	SDXp	SDXp	SRX'	SRX	OX
g_Z	O	*	.	P	P	O	O	O
gPrime_Z	P	*	P	.	O	O	O	O
gPrime_Y	P	*	P	O	.	O	O	O
g_X	O	*	O	O	O	.	O	O
g_W	O	*	O	O	O	O	.	O
g_V	O	*	O	O	O	O	O	.

Table 5.3: Homology table, showing orthologs, paralogs, and xenologs for the example from Figure 5.7. In this example, `xenolog_gene` has been reconciled with `xenolog_species` using the DTL event model (**Infer Transfers ON**). See Table 5.1 for a key to abbreviations. Notice that this table is not symmetric for xenologs. Xenolog pairs that are not defined (i.e., $(g, \widehat{g})$) are labeled with a star. Cells at the intersection of the column and row representing the same gene are labeled with a dot.

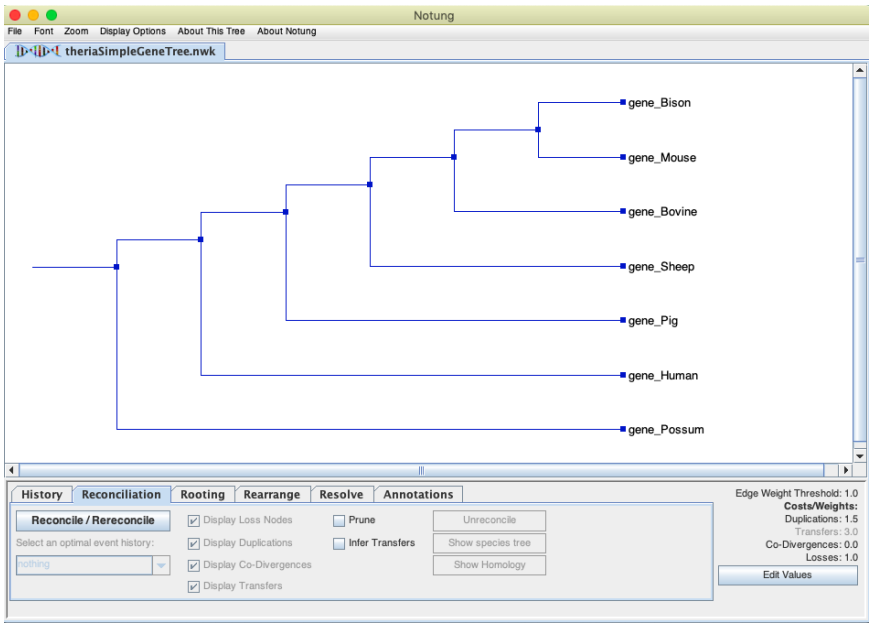


(a)

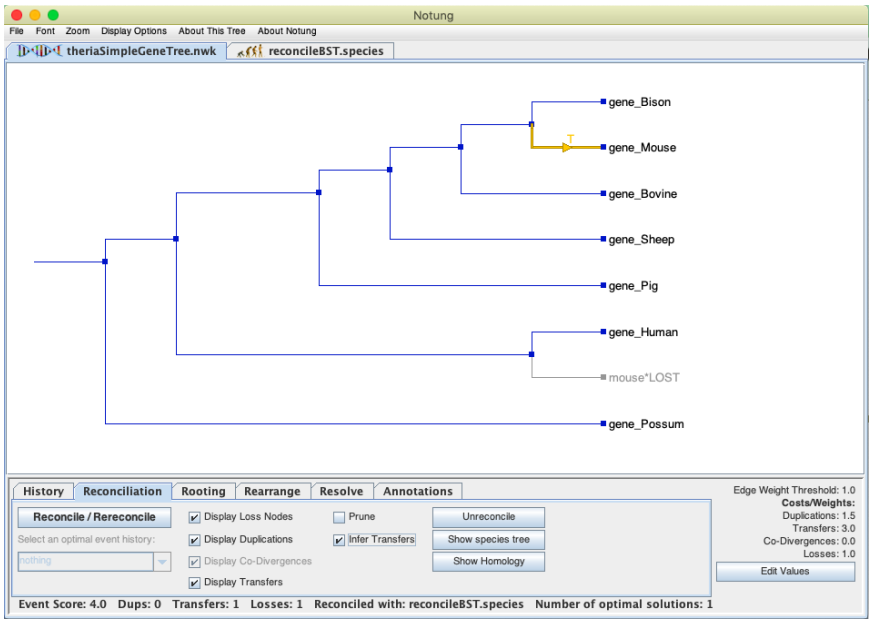


(b)

Figure 5.1: Reconciliation of the binary gene tree with a binary species tree in Figure 3.1(b). (a) The dialog box allows you to select which species tree to use, as well as the location of species labels in the gene tree. (b) The gene tree decorated with duplications (red boxes and D's) and losses (grey).

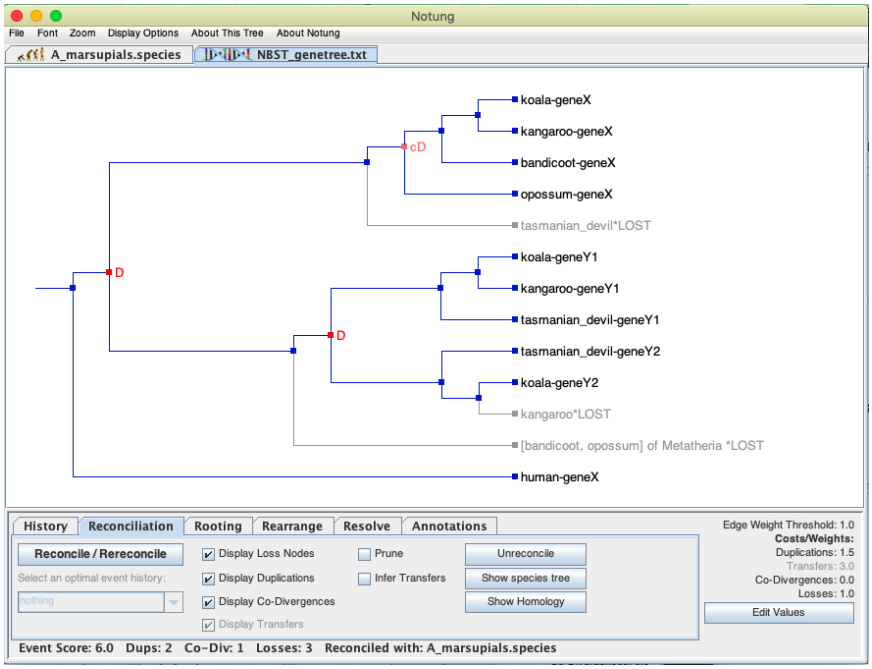


(a)

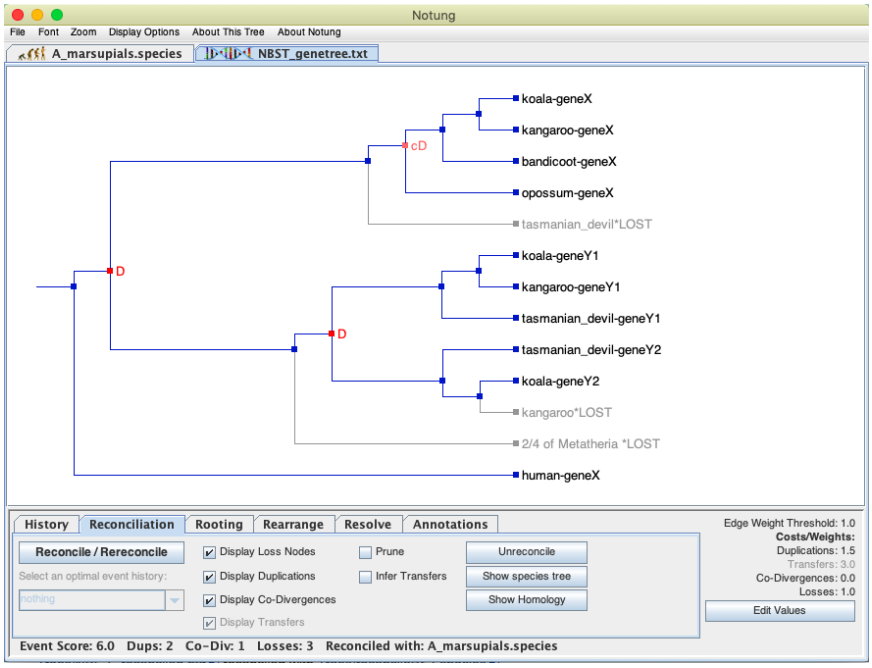


(b)

Figure 5.2: A binary gene tree reconciled with a binary species tree using the DTL-model. There are two reconciliations with the same, minimal Event Score. Both histories can be obtained by clicking on the green circle.

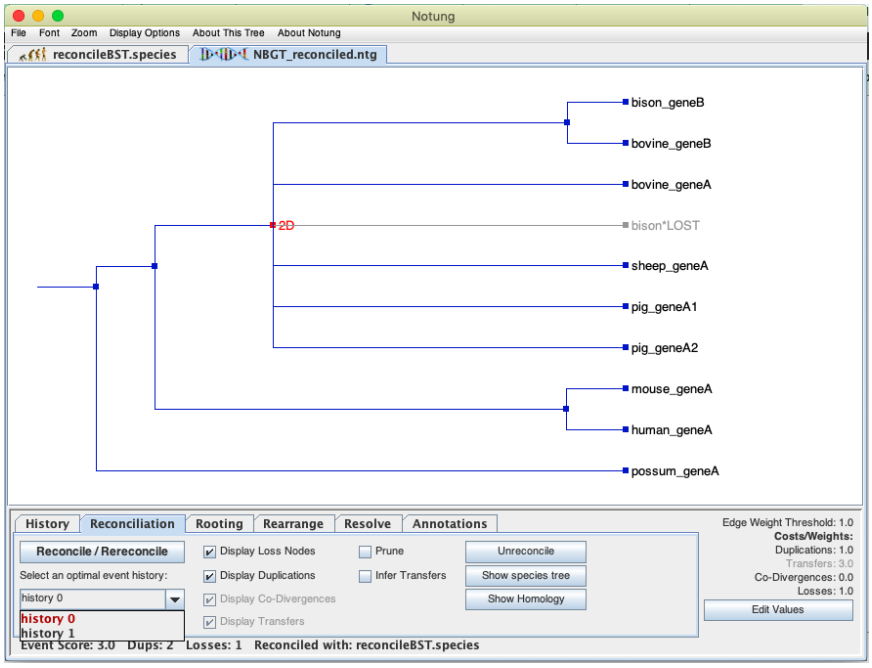


(a)

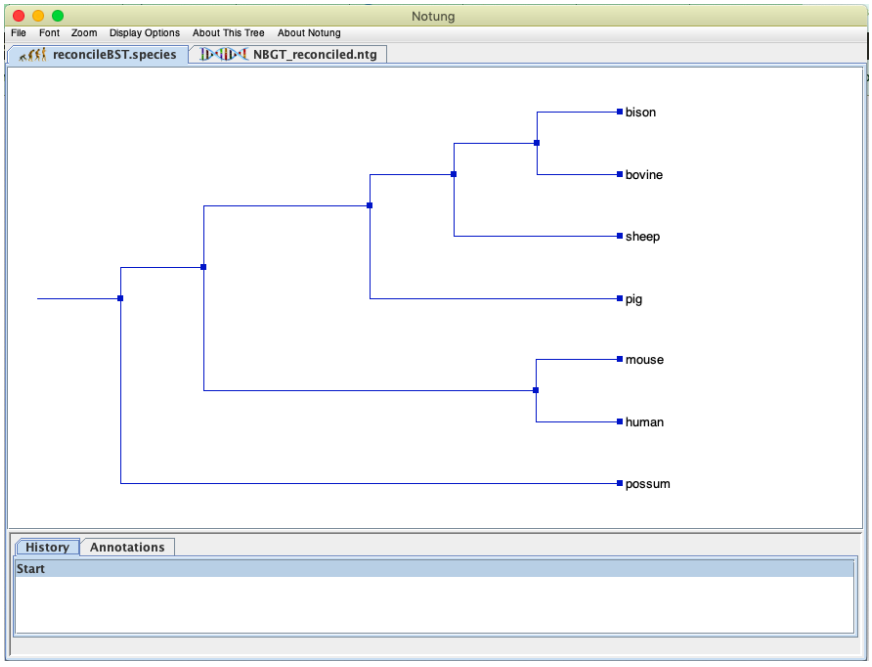


(b)

Figure 5.3: A binary gene tree reconciled with the non-binary species tree in Figure 4.1. Co-divergences are marked by pink cD's, while duplications are indicated with red D's, and transfers are indicated by yellow T's. Polytomy losses are labeled with the name of the associated polytomy, as well as either (a), the names of the species from which they are absent or (b) the number of species from which they are absent.



(a)



(b)

Figure 5.4: Reconciliation of (a) a non-binary gene tree with the binary species tree in (b). More than one duplication may be inferred at polytomies in the gene tree. In addition, it is possible to have more than one optimal event history, as seen in the lower left-hand corner of the reconciliation panel in (a).

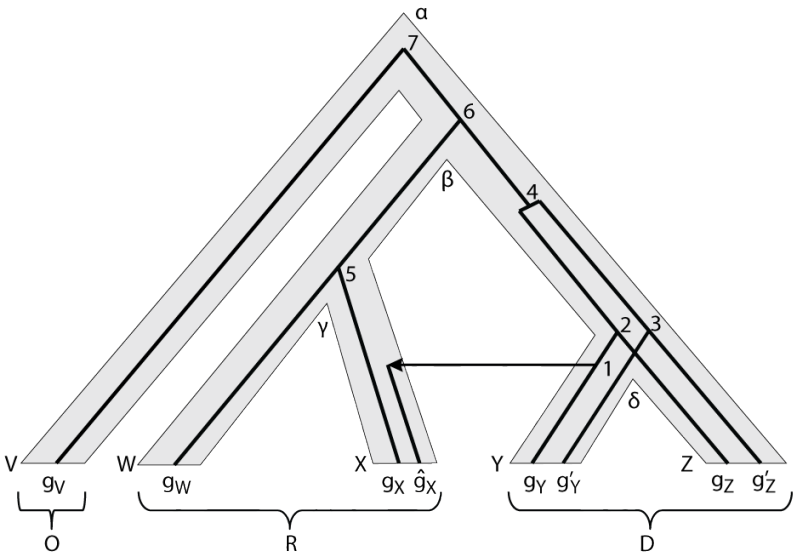


Figure 5.5: *Xenolog classes* of a gene family evolving through speciation, duplication, and horizontal gene transfer. The gene tree (thin black lines) is shown embedded in the species tree. In the history of this hypothetical family, two events occurred: a duplication (gene node 4) in species δ and a transfer (black arrow) from species Y to species X. The cenancestor of the transfer is species β. Species that are more closely related to the donor and recipient species are indicated by sets D and R, respectively. Species that are equally related to both the donor and recipient species are indicated by the set O. Notung's homology classification for this example is shown in Figure 5.7 and Table 5.3.

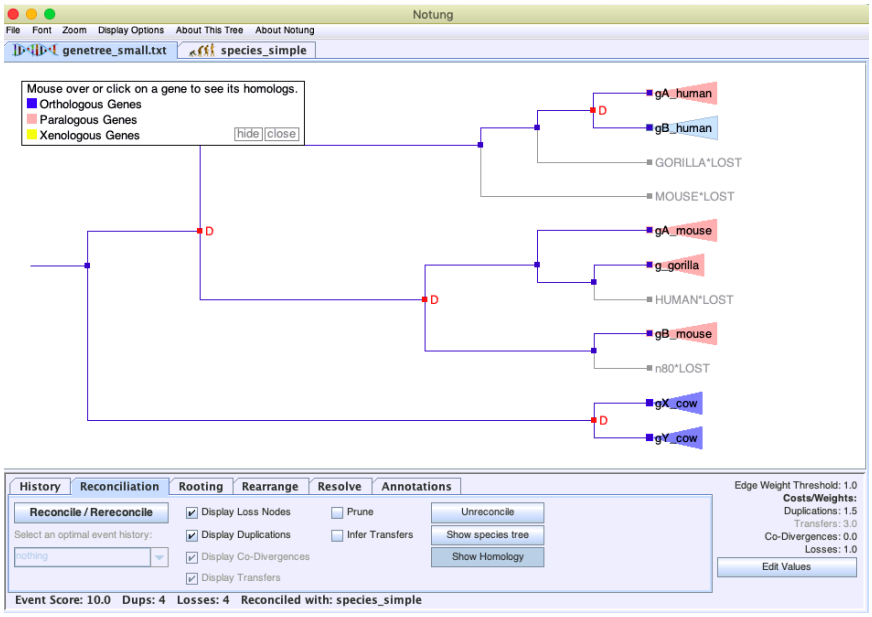


Figure 5.6: Notung's point and click interface showing the homology relationships (orthologs and paralogs) to gene `gA_human` (highlighted in light blue). Orthologs are highlighted in dark blue, paralogs in pink. Gene tree `genetree_SMALL` has been reconciled with `speciestree_SMALL` using the DL event model (**Infer Transfers OFF**).

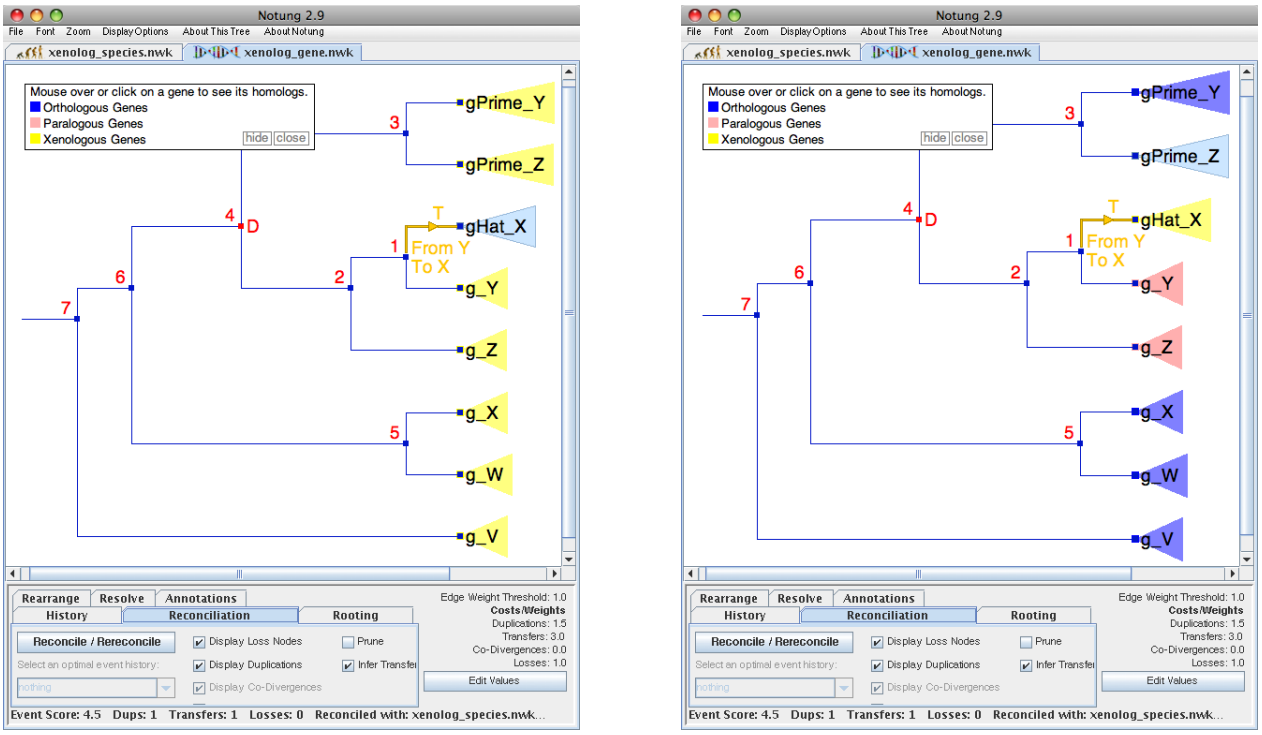


Figure 5.7: Notung's point and click interface showing the homology relationships (orthologs, paralogs, and xenologs) in gene tree `xenolog_gene`, which has been reconciled with `xenolog_species` using the DTL event model (**Infer Transfers ON**). This reconciliation has two events: a duplication (node 4) and a transfer (node 1). Shown are the homology relationships for genes (*left*) `gHat_X` and (*right*) `gPrime_Z`, both highlighted in light blue. Orthologs are highlighted in dark blue, paralogs in pink, and xenologs in yellow. Since gene `gHat_X` is a recipient of the transfer, all other genes are xenologs to `gHat_X` and are highlighted in yellow. The different xenolog subclassifications are not indicated.

Chapter 6

Rooting Mode

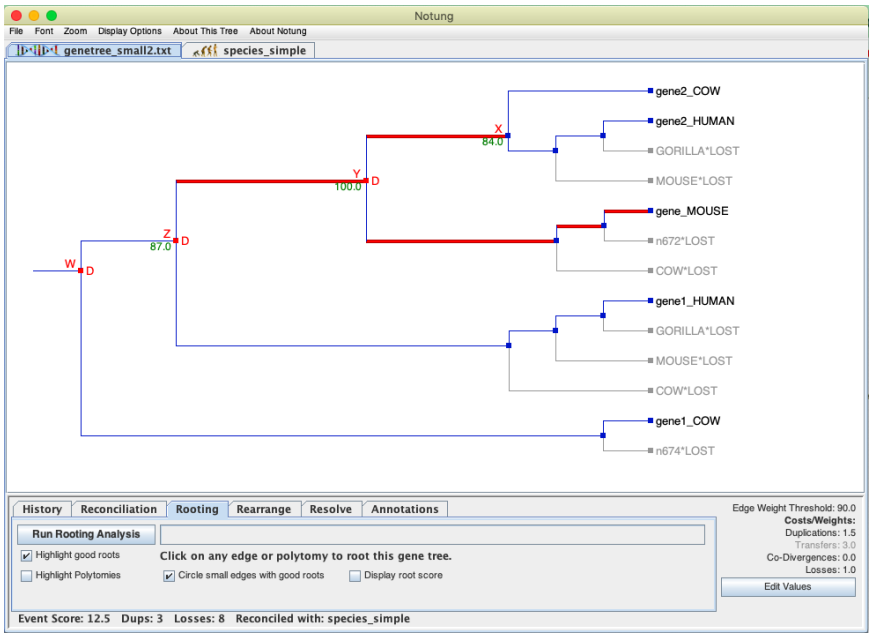
In Rooting mode, the event parsimony can be used to infer the root of a gene tree. Notung's Rooting Analysis calculates a **root score** for each edge in the tree, corresponding to the DTL Score of the tree rooted on that edge. Note that the Rooting Analysis computes root scores, but does not change the tree. The user must root the tree explicitly by clicking on an edge in the tree panel. Rooting mode can also be used to root a tree manually by clicking on any edge in the tree at any time, even if the Rooting Analysis has not been performed.

When the Rooting Analysis is complete, edges with the minimum root score are highlighted in red. Notung also highlights edges with near-optimal scores in pink. Edges with scores that are greater than the minimum by at most 5 percent of the difference between the maximum and minimum score are highlighted in pink. Figure 6.1(a) shows the gene tree from Figure 5.1 after the Rooting Analysis has been applied. Note that optimal rooting edges are highlighted in red, but the gene tree topology is unchanged from Figure 5.1. Figure 6.1(b) shows the tree after it has been rooted by clicking in the tree panel.

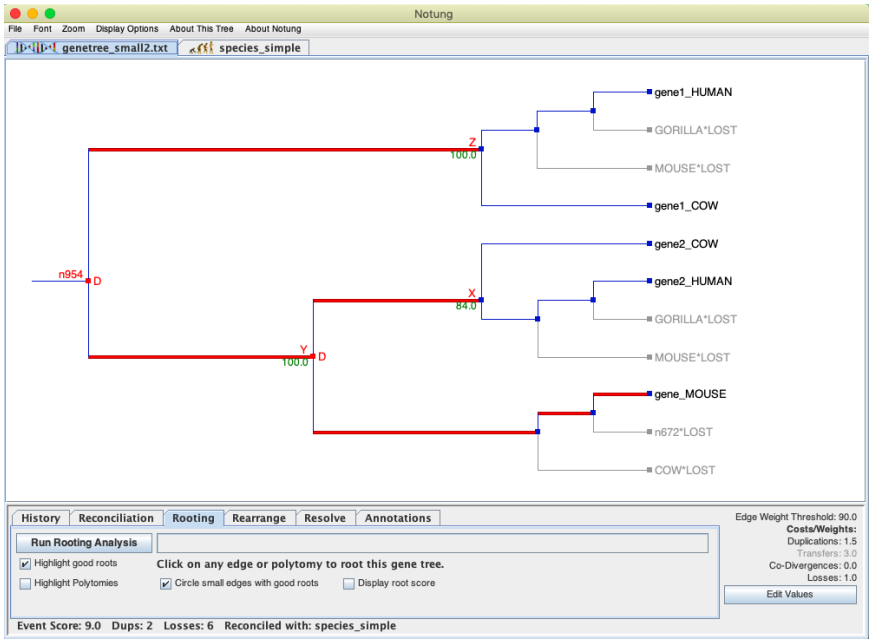
When the *gene tree is binary*, applying Notung's Rooting function with DL model results in a binary gene tree with a root score on each edge. This score is a weighted sum of the number of duplications, co-divergences, and losses. This score also represents the optimal events cost when the binary gene tree is rooted at that particular branch. By default, the cost of co-divergences is set to zero. Co-divergences will only influence the root score if this cost is explicitly set to a positive number by the user. For more information on setting parameters, see Chapter 3.5 - Parameter Values on page 22.

Rooting analysis with transfers proceeds somewhat differently from rooting with the DL-model because temporal feasibility must be considered.

Rooting analysis when the *gene tree is non-binary* is currently only available for DL-model



(a)



(b)

Figure 6.1: (a) The gene tree from Figure 5.1 after completing the Rooting Analysis. (b) The rerooted tree, after the user has clicked on an edge to designate the root.

(without transfers). Rooting analysis with a non-binary gene tree differs from the binary case in that root scores are assigned to polytomies, as well as edges. Edges and polytomies in the original tree are assigned the DL Score associated with rooting on that edge or polytomy. If

rooting on a polytomy in a non-binary gene tree produces the minimum or near-minimum score, that node will be circled and the vertical edge representing that polytomy will be highlighted in the appropriate color (Figure 6.2).

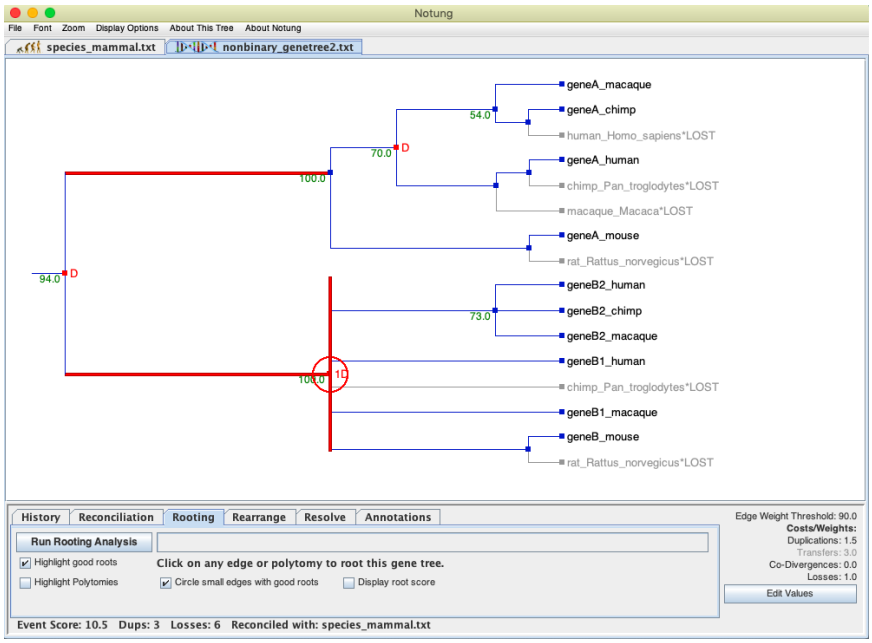


Figure 6.2: Rooting analysis for a non-binary gene tree. The optimal root locations are colored in red. If an edge represented by the polytomy can be selected as an optimal root, the polytomy will be circled and colored in red.

To re-root the tree, click on any edge or polytomy in the tree panel. You may root the tree on any edge, not just the highlighted edges. Notung will root the tree on that edge (or polytomy), and recalculate the reconciliation. The DL Score of the new, rooted tree is displayed in the bottom-left corner of the screen.

Some tree reconstruction programs represent an unrooted tree as a rooted tree with a trifurcation (a polytomy with three children) at the root. Notung cannot distinguish between an unrooted, binary gene tree and a rooted gene tree that has a single trifurcation. If such a gene tree is opened and reconciled in Notung, a notification will appear to inform the user that this tree may, in fact, be an unrooted, binary gene tree. Notung will assume that the tree is rooted and non-binary, and will draw the tree and issue diagnostic messages accordingly. If you want Notung to treat the tree as an unrooted binary tree, the trifurcation can be removed by rooting the tree in the **Rooting panel**. Clicking on any edge will convert the gene tree into a rooted binary tree. To select the edge with the optimal DL Score, run the **Rooting Analysis** first.

> *NOTE*: If the tree has not been reconciled before running a Rooting Analysis, Notung will reconcile it automatically. You will be asked to select a species tree

for reconciliation (see Chapter 5 - Reconciliation Mode on page 40).

6.0.1 Rooting with Transfers

In Rooting mode with transfers, Notung first visits each edge in the tree in turn, infers the *candidate* optimal reconciliation(s) of the tree rooted on that edge and determines the associated DTL score. This is the provisional root score for that edge, but is only a valid root score if at least one of the candidate optimal reconciliations is also feasible. Once the provisional root scores have been determined, the candidate reconciliations for the edge(s) with the lowest provisional root score are tested for temporal feasibility until a feasible solution is found. To reduce the running time, the remaining candidate reconciliations are not tested, since the existence of one temporally feasible, optimal reconciliation is sufficient to determine that the edge is a feasible root with minimum DTL score. If none of the candidate reconciliations associated with an edge are feasible, then that edge is not a viable root.

In the GUI, Notung reports all edges with a minimum root score that have at least one feasible solution, in contrast to the DL model, where Notung reports the root score for every edge in the tree. With transfers, only edges corresponding to minimum cost roots are reported in order to reduce the computational burden of testing (possibly many) candidate solutions for feasibility. It is possible to infer the cost of sub-optimal roots from the command line. Please refer to Section 12.7 - Rooting trees from the command line with the DTL model for details.

In the GUI, edges with the lowest root score are highlighted in red in the rooting panel, once the rooting analysis is complete. To re-root the tree, click on one of the red edges. (You can re-root the tree on any edge, but only red edges will give a minimum cost reconciliation.) To see the reconciliation(s) associated with that root, go to the reconciliation panel and re-reconcile the tree.

If no minimum cost edge has a feasible solution, then Notung does not report a root. Instead, a pop-up warning will be displayed.

To find optimal root edges:

1. Click the **Rooting** tab to enter **Rooting** mode.
2. Click "**Run Rooting Analysis**."

Good roots will be highlighted. If highlighted edges are small, they are circled in the appropriate color to help the user locate them visually. Use the Zoom feature (see Chapter 11.2 - Zoom on page 95) to zoom in on these edges.

To show/hide Rooting Analysis results:

In Rooting Mode, the task panel contains several check boxes that allow the user to specify what rooting related information should be displayed.

- Click the appropriate checkbox in the **Rooting** task panel.

  √ **Highlight good roots (default: ON)** When the "**Highlight good roots**" box is checked, the low-scoring root edges in the tree are colored (red for the minimum score, pink for scores within 5 percent of the minimum). To remove the highlighting, uncheck the box.

  √ **Highlight polytomies (default: OFF)** When the "**Highlight polytomies**" box is checked, vertical edges representing polytomies in the gene tree, if any exist, are highlighted and circled in cyan.

  √ **Circle small edges (default: ON)** When the "**Circle small edges**" box is checked, appropriately colored circles appear around small, low-scoring root edges in the tree, if any exist, making it easier to locate them in large trees. To remove the circles, uncheck the box.

  √ **Display root score (default: OFF)** When the "**Display root score**" box is checked, the associated D/L Score appears next to each edge and polytomy in the tree. Rooting scores are shown in pink to distinguish them from edge weights, which are displayed in green.

  NOTE: For efficient run times, Notung only tests optimal roots for temporal feasibility. Therefore, rooting with transfers cannot display root scores for each edge, as they may not be feasible.

To re-root the tree:

- Click on any edge or polytomy of the tree in the tree panel.

Chapter 7

Rearrange Mode

Weakly-supported edges, as indicated by low edge weights, often imply that the inferred history associated with those edges may not be accurate. Notung can rearrange weakly-supported regions in a gene tree to produce alternate event histories with minimum DL Score. When these edges or regions are rearranged, the structure of strongly-supported edges or regions stays intact. Any edge that is added as a result of rearrangement will be not be assigned an edge weight. Since support for edges is determined by edge weight, Notung's rearrangement function requires that the gene tree include edge weights which assess how well each edge is supported by sequence data. These edge weights can be bootstrap values, probabilities, or branch lengths.

> *NOTE*: Currently, Notung can not perform rearrangement with non-binary species trees.

Weak edges are defined as those edges with weights below the Edge Weight Threshold. Selecting the "**Highlight weak edges**" checkbox in **Rearrange** mode will highlight all weak edges in yellow, allowing the user to see which edges will be considered for rearrangement (see Figure 7.1). This option is only available in **Rearrange** mode. The yellow highlighting will disappear when another mode is selected. As a default, the Edge Weight Threshold is 90% of the maximum edge weight. While this is a good starting place for bootstrap values, it may not be appropriate for probabilities or branch lengths. The threshold can be adjusted by the user; see Chapter 3.5 - Parameter Values on page 22 for information on how to change the Edge Weight Threshold. Notung also considers any edge without an assigned weight to be a weak edge. If Notung's rearrangement function is applied to a tree with no edge weights, it will consider all edges to be weak, and will find all trees that are optimal when only events are considered (*i.e.* those trees with a minimal Event Score).

The **Rearrangement** function can be applied to a non-binary gene tree when the species tree is binary (Figure 7.2). Notung will replace each polytomy with an arbitrary binary resolution, inserting new nodes and edges. These new edges are treated as weak edges. The standard rearrangement algorithm is then applied to the resulting binary tree to determine the rearrangement that results in a minimal Event Score. Note that it is immaterial how the polytomies are initially resolved, because subsequent rearrangement will result in a minimum cost tree. Rearrangement cannot be performed when the species tree is non-binary.

7.1 Alternate Optimal Hypotheses

When rearranging a gene tree, there may be more than one tree that (1) agrees with the original tree at strongly supported edges and (2) has minimal Event Score. If there are many such trees, considering all of them may be a daunting task. Notung addresses this issue by partitioning the set of all optimal trees into subsets in such a way that any tree in a given subset can be generated from any other tree in the subset by a series of node interchanges.

All trees in any given subset are instances of the same event history. An **event history** describes a series of events (duplications and losses) and the location in the species tree where they occurred. "A duplication in the common tetrapod ancestor, a loss in the fish lineage and three duplications in mouse" is an example of an event history. To see that more than one tree can have the same event history, note that "three duplications in mouse" corresponds to the subtree `((g1_mouse, g2_mouse), (g3_mouse, g4_mouse))`, as well as the subtree `(((g1_mouse, g2_mouse), g3_mouse), g4_mouse)`.

If multiple minimum cost trees are found, Notung presents one tree from each subset (*i.e.* one representative of each event history) to the user and provides a point and click interface that allows the user to inspect any other tree in that subset. Initially, Notung arbitrarily selects one event history to present in the tree panel. The other optimal histories may be viewed using the drop-down menu labeled "Select an optimal event history," which gives a list of up to 50 optimal event histories. The user can perform Same Cost Swaps on a tree to explore the space of all optimal trees corresponding to the current event history. **Same Cost Swaps** are node interchanges that result in another tree with an optimal DL score. Clicking the "**Examine same-cost swaps**" button will highlight all **swappable nodes**, nodes that can be manually swapped without changing the DL Score.

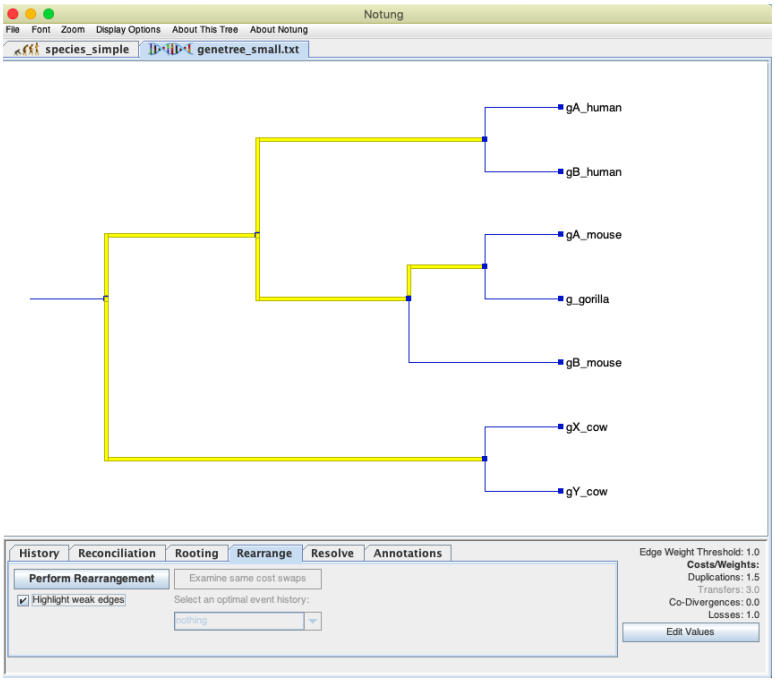
If there are more than 50 optimal event histories, they can be generated using the Command Line Interface (see Chapter 12 - Command Line Options and Batch Processing on page 98). Note that both the drop down menu and command line options give distinct optimal event histories, but for the DL event model, *do not generate all optimal gene tree rearrangements.* It is only possible to view all trees by performing same cost swaps using the point and click

interface in the GUI.

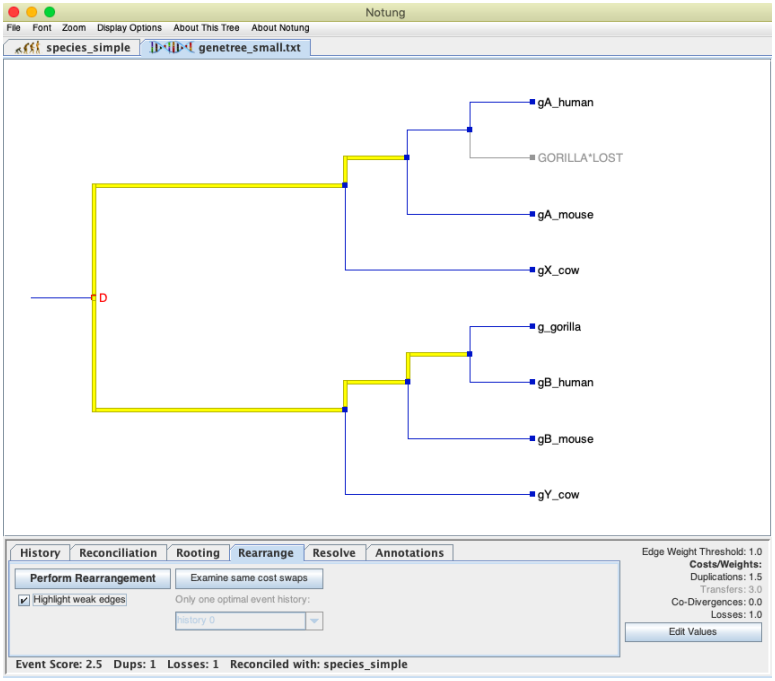
For further details on Notung's rearrangement algorithm see:

D. Durand, B. V. Halldorsson, B. Vernot. A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. Journal of Computational Biology, 13(2): 320-335, 2006.

H. Lai, M. Stolzer, and D. Durand. Fast Heuristics for Resolving Weakly Supported Branches Using Duplication, Transfers, and Losses. Proceedings for RECOMB International Workshop on Comparative Genomics: 298-320, 2017.

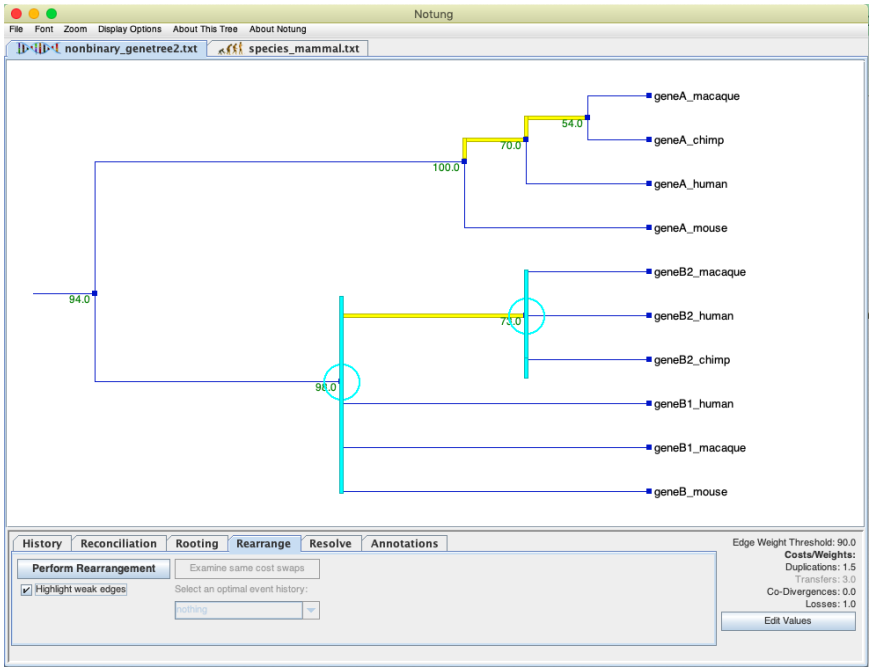


(a)

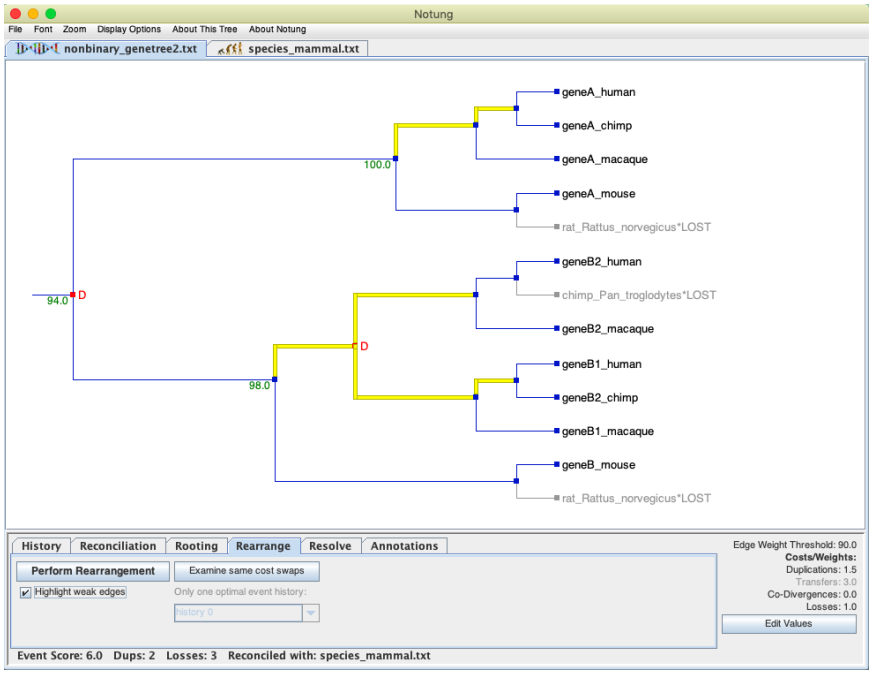


(b)

Figure 7.1: (a) The gene tree from Figure 5.1 with weak edges highlighted. (b) After clicking "Perform Rearrangement," the rearranged tree appears in the tree panel. Weak edges are still highlighted in yellow.



(a)



(b)

Figure 7.2: (a) When rearranging the non-binary gene tree, weak edges are highlighted in yellow. These edges, as well as the polytomies, highlighted in cyan, will be rearranged to produce the binary tree with the minimal D/L Score. (b) After the tree is rearranged, weak edges are highlighted in yellow. Notice that new edges have no edge weight and are considered weak.

7.2 Rearrangement Commands

NOTE: Currently, Notung can only perform rearrangement with transfers via the command-line interface. See Section 12.8 - Non-binary gene trees with the DTL model on page 117 for more information.

To rearrange the gene tree:

1. Click the **Rearrange** tab to enter **Rearrange** mode.
2. Click "**Perform Rearrangement**."

   A minimum cost rearrangement tree will appear in the tree panel as shown in Figure 7.1(b). Note that weak edges, highlighted in yellow, will not have edge weights. Some or all of these are edges that do not correspond to any bipartition (split) represented in the original tree. The appropriate weights for these edges are not known.

   NOTE: If asked to rearrange a tree that has not been reconciled, Notung will reconcile it automatically. In this case, the user is asked to select a species tree for reconciliation.

To highlight all weak edges (default: OFF):

- Click the "**Highlight weak edges**" checkbox.

  All weak edges in the tree will be highlighted in yellow.

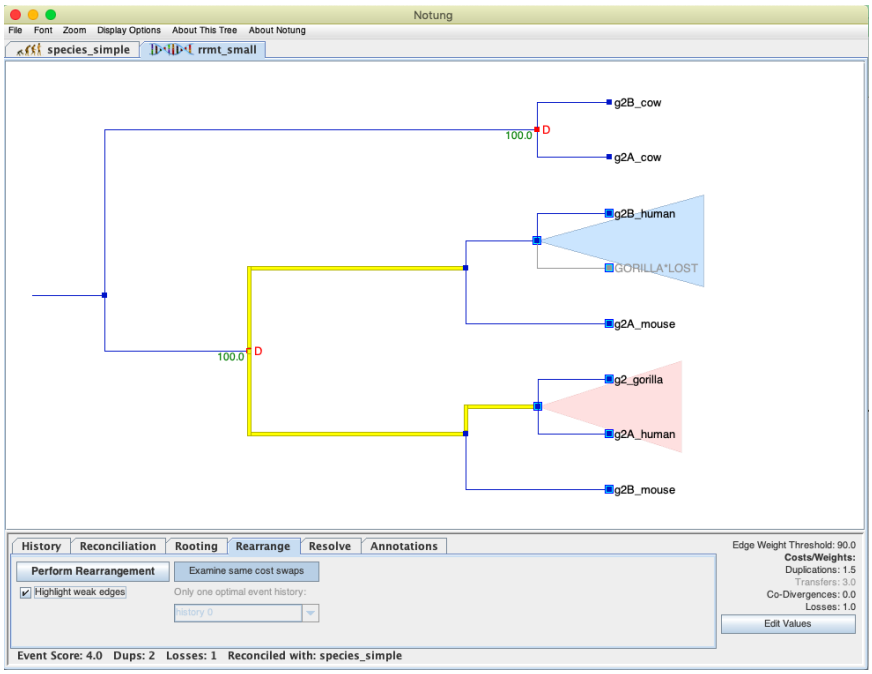
To view alternate optimal event histories:

If more than one optimal event history exists for a rearranged tree, the drop down menu "**Select an optimal event history**'' will be enabled.

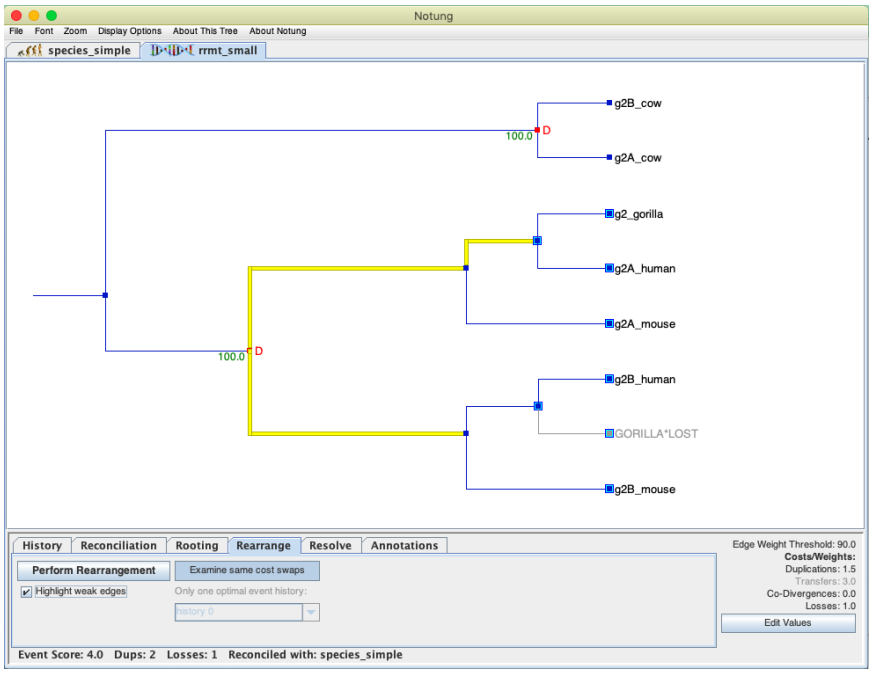
- From the drop-down menu, select an alternate event history.

  The tree panel will now show a new tree corresponding to the selected alternate history.

If there is only one optimal history or the tree has not yet been rearranged, the drop down menu will be grayed out.



(a)



(b)

Figure 7.3: When examining same cost swaps, swappable nodes are marked with the enlarged square. (a) Same cost swaps are highlighted. The selected node, shown in blue, can be swapped with the node highlighted in orange. Clicking first on the blue node and then on the orange node results in the alternate optimal tree shown in (b).

To swap individual nodes:

1. Click the "**Examine same cost swaps**" button in the right column on the **Rearrange** task panel.

   NOTE: If there are no swappable nodes in the tree or if the tree has not yet been rearranged, this button will be grayed out.

   Swappable nodes are marked with an enlarged blue and cyan square. As you pass the mouse over a swappable node it will be highlighted with a blue triangle. Other nodes that can be interchanged with it are temporarily highlighted with a light orange triangle, as shown in Figure 7.3(a). If you have zoomed in, some swappable nodes may be outside the boundaries of the tree panel. Swappable nodes that are not currently visible are indicated by arrows in the tree panel, pointing in the direction of those nodes. These can be seen by scrolling in the direction of the arrow.

2. Click a node to swap. The node you selected is highlighted with a blue triangle. Nodes with which it can be swapped are now highlighted with red triangles.

3. Click a second node to complete the swap (see Figure 7.3(b)).

NOTE: When a user selects a different alternate event history from the "**Select an optimal event history**" list, Notung rebuilds the tree from data saved at the time of rearrangement. Any manual swaps made to a previously viewed event history will be lost. Therefore, if you wish to save information after a manual swap, you must save your tree. See Chapter 3.3 - Opening and Saving Trees on page 15 for more information.

Chapter 8

Resolve Mode

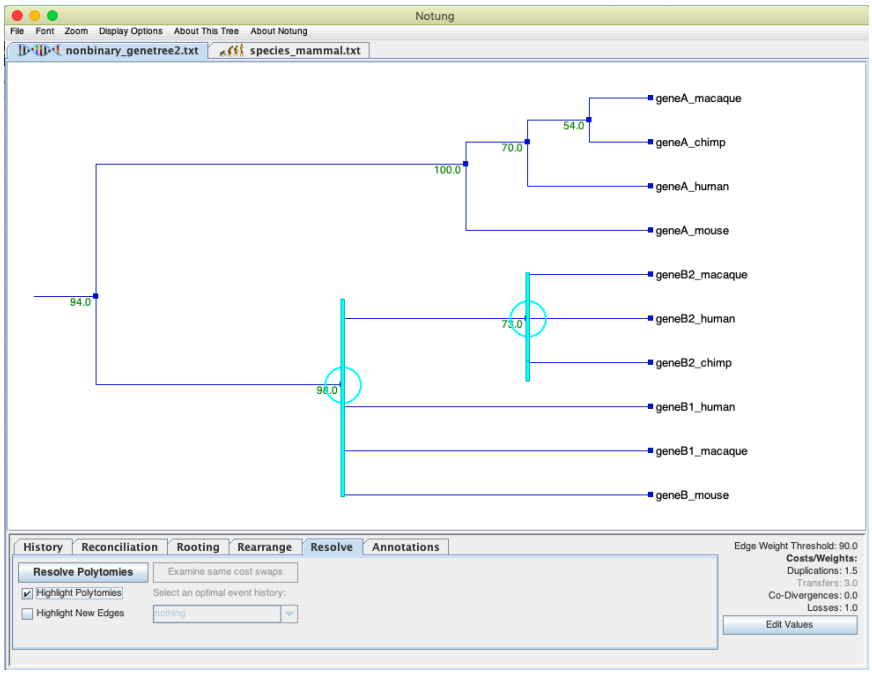
Resolve mode is only applicable to non-binary gene trees. Its function is to resolve polytomies in a non-binary gene tree by comparing it with a binary species tree, resulting in one or more binary tree(s) with minimal DL Score.

> *NOTE*: Notung currently cannot reconcile non-binary gene trees using the transfer algorithm. Therefore, such trees cannot be resolved with transfers.

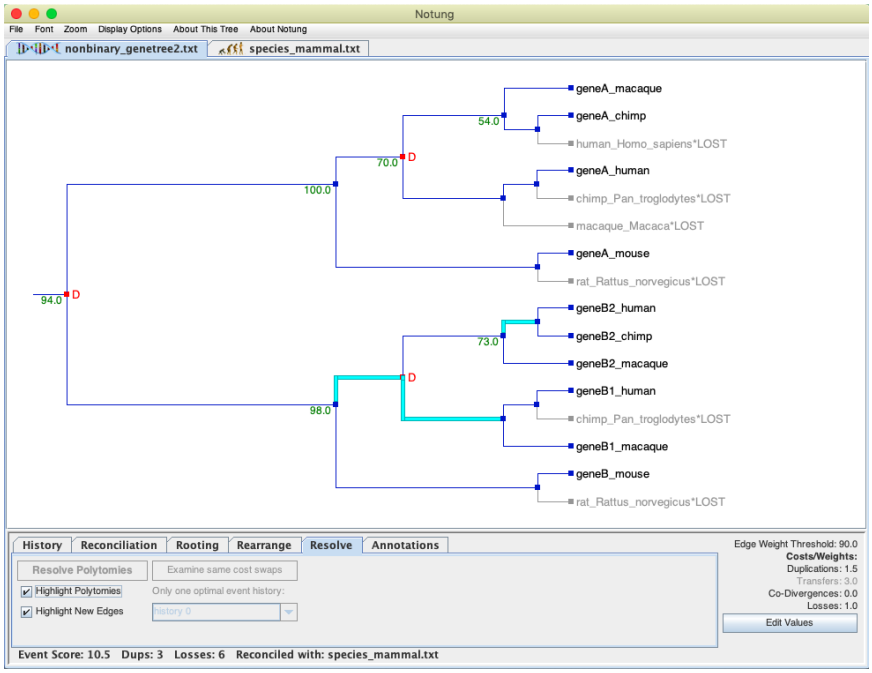
Specifically, the **Resolve** function removes all polytomies in the original gene tree, and uses an algorithm similar to the rearrangement algorithm to replace them with new edges such that: 1) the new tree is binary, and 2) the new tree has optimal DL Score. Note that each edge in the original non-binary gene tree still exists in the resulting binary gene tree.

There may be more than one binary tree that agrees with the input tree at all edges except polytomies and has minimal DL Score. In this case, the user can investigate these optimal alternate hypotheses using a point and click interface as in **Rearrange** mode. See Section 7.1 - Alternate Optimal Hypotheses on page 74 for a more detailed explanation of alternate hypotheses.

Selecting the "**Highlight Polytomies**" checkbox will highlight in cyan all vertical edges representing polytomies in the tree, allowing the user to see which nodes will be resolved. After running the resolve algorithm, the "**Highlight New Edges**" checkbox will be selected, and will highlight in cyan all those edges in the gene tree that were previously represented by the polytomy (see Figure 8.1). This option is only available in **Resolve** mode.



(a)



(b)

Figure 8.1: (a) Polytomies in the gene tree can be highlighted in cyan while in the Resolve task mode. (b) After the polytomies are resolved, edges that were not present in the original tree are highlighted in cyan.

To resolve the gene tree:

1. Click the **Resolve** tab to enter **Resolve** mode.
2. Click "**Resolve Polytomies**."

   A minimum cost binary resolution of all polytomies in the tree will appear in the tree panel. Note that the new edges will not have edge weights.

   If the gene tree is binary, the "**Resolve Polytomies**" button will be grayed out.

NOTE: If asked to resolve a tree that has not been reconciled, Notung will first invoke the reconciliation algorithm. In this case, the user is asked to select a species tree for reconciliation.

To highlight all polytomies (default: OFF):

- Click the "**Highlight Polytomies**" checkbox.

  All vertical edges representing polytomies in the tree will be highlighted.

To highlight all new edges (default: ON, after resolving):

- Click the "**Highlight New Edges**" checkbox.

  All edges that were represented by the polytomies in the original tree will be highlighted.

To view alternate optimal event histories:

1. If more than one optimal event history exists for a resolved tree, the drop down menu "**Select an optimal event history**" will be enabled.
2. From the drop-down menu, select an alternate event history.

   The tree panel will now show a new tree corresponding to the selected alternate history.

If there is only one optimal history or if the polytomies have not been resolved, the drop down menu will be grayed out.

To swap individual nodes:

1. Click the "**Examine same cost swaps**" button on the **Resolve** task panel.

   > *NOTE:* If there are no swappable nodes in the tree or if the polytomies have not been resolved, this button will be grayed out.

   Swappable nodes are marked with an enlarged blue and cyan square. As you pass the mouse over a swappable node, other nodes that can be interchanged with it with are temporarily highlighted with a light orange triangle. Swappable nodes that are not currently visible in the tree panel (for instance, if you have zoomed in), are indicated by arrows in the tree panel pointing in the direction of those nodes.

2. Click a node to swap.

   The node you selected is highlighted with a blue triangle. Nodes with which it can be swapped are now highlighted with pink triangles.

3. Click a second node to complete the swap.

   NOTE: When a different alternate event history is selected in the "Select an optimal event history" list, Notung rebuilds the tree from data saved at the time of resolution. Any manual swaps made to a previously viewed event history will be lost. Therefore, if you wish to save information after a manual swap, you must save your tree. See Chapter 3.3 - Opening and Saving Trees on page 15 for more information.

Chapter 9

History

The state of a gene tree changes each time a Notung operation, such as rooting, rearrangement, reconciliation, or resolution, is performed on the tree. Notung maintains a history of state changes for each gene tree. This history can be accessed via the History panel, allowing the user to return to and operate on a previous state, or visually compare the state before and after a task is performed.

Notung lists the states in the history panel by task name (see Figure 9.1). The first entry in the list is always **Start**, which is the state of the tree when loaded; others entries may include **Changed Parameter Values, Reconciled, Rooting Analysis, Rooting on X, Notung Rearrange, Notung Resolve Polytomies, Select Alternate Optimal History,** and **Swapped Y and Z**, where X is an edge and Y and Z are swapped nodes. The list proceeds from top to bottom in the order tasks were performed, and includes the Event Score for each state.

NOTE: Previous states in the History panel are not saved in a file. When the gene tree file is closed, the history associated with the current tree is lost. To save trees associated with intermediate states, select the state and click "File → Save As."

NOTE: Parameter values are saved with each state in the history. For each state in the history, the parameters will correspond to those values used at the time the operation was performed. Any subsequent changes to parameter values will not be applied retroactively.

To view previous states of the gene tree:

1. Click the History tab to enter History mode.
2. Click on an item in the list.

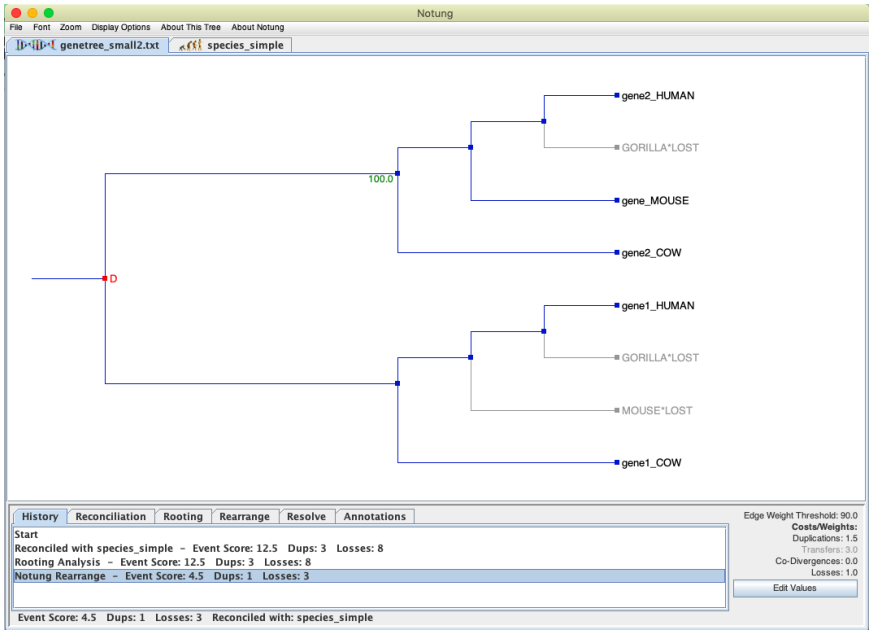


Figure 9.1: The history of a gene tree that had been reconciled, rooted, and rearranged. Currently, the state of the tree after reconciliation and prior to rooting is selected and displayed in the panel.

Chapter 10

Annotations

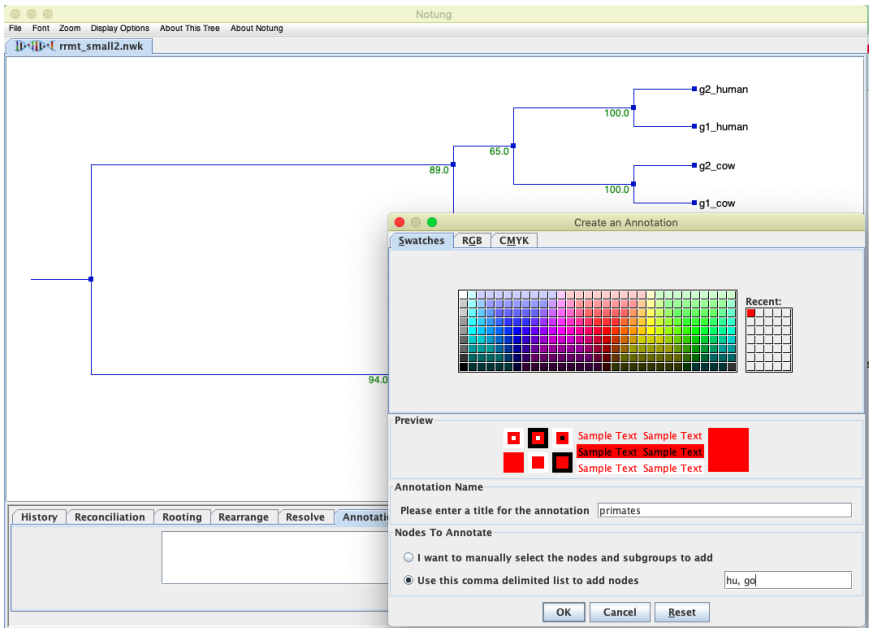
Notung can annotate the leaf nodes of both gene and species trees with colors specified by the user. For example, the annotation function can be used to color all nodes associated with a particular taxonomic group (*e.g.*, plants) or a particular subfamily (*e.g.*, HSP70). This can help visually differentiate gene clusters in a large and complex tree, or highlight related nodes that are distantly located in a tree.

The "New" button in the Annotations task panel opens the annotations dialog window (see Figure 10.1), where the user can set the annotation parameters. Each annotation consists of a title used to identify it, a color, and a specification of the nodes that are included in the annotation. The title of an annotation is simply an alphanumeric string used to distinguish it. You may use any string of characters as long as it is unique. The set of nodes associated with a given annotation can be specified in two ways, by pattern matching or by selecting them manually. In the first case, the user provides one or more alphanumeric strings, which are compared with all leaf node names. Leaf nodes that contain one or more of the specified strings as a substring are added to the annotation. Alternatively, nodes can be manually added to the annotation by clicking on them.

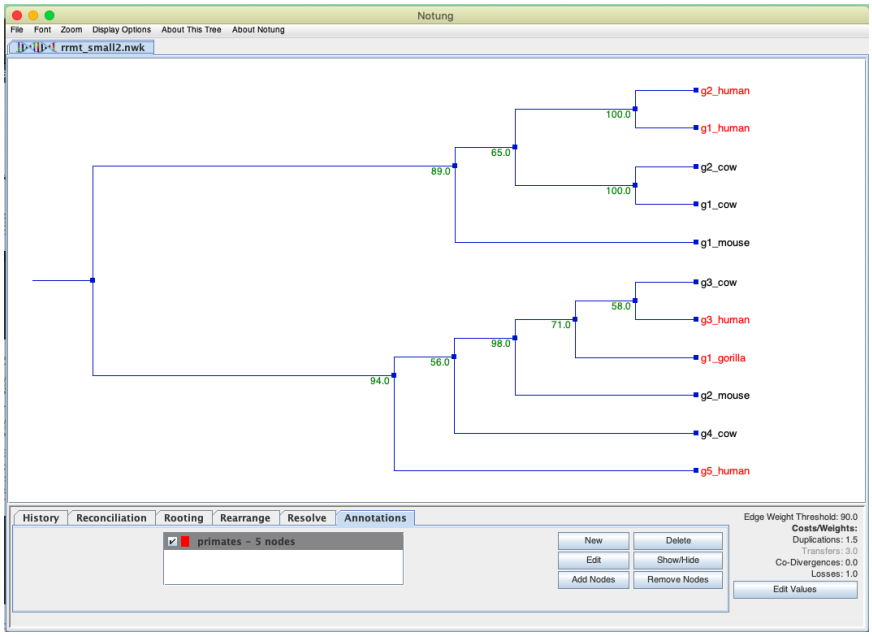
All annotations for the currently selected tree are shown in the list box in the Annotations task panel (see Figure 10.1(b)). After an annotation is created, individual nodes can be added to it or removed from it, manually. Annotations can be edited to modify the list of pattern matching terms or to change the color associated with the annotation. Annotations can be shown or hidden at any time.

NOTE: A single node can match more than one annotation, but will only be colored by the most recently created annotation.

NOTE: Annotations only apply to the tree that is currently selected, but can be exported and



(a)



(b)

Figure 10.1: (a) The annotations dialog box. This figure shows the creation of an annotation, with the title "primates", associated with the color red, and the pattern matching terms "hu" and "go". (b) A fully annotated gene tree. In the Annotations task panel, the list box shows the annotations associated with the currently selected tree. A check box indicates whether the annotation is hidden (unchecked) or showing (checked). The number next to each annotation refers to the number of leaf nodes currently colored by that annotation. If the annotation is hidden, this number is zero.

then imported into another tree. See subsections on importing and exporting annotations at the end of this section.

To create an annotation using pattern matching (recommended):

1. Click the **Annotations** tab to enter **Annotations** mode.
2. Click the "**New**" button in the task panel. A dialog box appears.
3. Select a color in the color palette for the annotation. If you do not select a color, Notung selects red by default.
4. Enter the title of the annotation in the text field in the center of the dialog box.
5. Select the radio button marked "**Use this comma-delimited list to add nodes**." (This button is selected by default.)
6. Enter the pattern matching term(s) in the text field, separated by commas. If no terms are entered, Notung will use the title of the annotation as a pattern matching term.

   For example, if you want to annotate all the node labels containing HU, enter "HU" in the text field. Notung will annotate any node with a label that contains "HU" as a substring, such as `g1_human` and `g2_human`. If you want to annotate all node labels containing "HU" and "GO," enter "Hu, Go" in the text field. This will also annotate the node `g1_gorilla`, as seen in Figure 10.1.

   NOTE: This process is not case sensitive.
7. Click "**OK**."

   Nodes with names that match a string in the comma-delimited list will change color (*e.g.*, Figure 10.1).

If a single node corresponds to more than one annotation, the node will be in the color dictated by the most recently added annotation. The newer annotation will continue to take precedence until the shared node is manually removed from that annotation, the annotation is hidden, or a new, conflicting annotation is added. For example, adding an annotation in yellow for "g1" would change the color of `g1_human`, `g1_cow`, `g1_mouse`, and `g1_gorilla` to yellow.

To create an annotation with manually added nodes:

1. Click the **Annotations** tab to enter **Annotations** mode.

2. Click the "**New**" button in the task panel. A dialog box appears.

3. Select a color in the color palette to use for the annotation (if you do not select a color, Notung selects red by default.)

4. Enter the title of the annotation in the text field in the center of the dialog box.

5. Select the radio button marked "**I want to manually select the nodes and subgroups to add**."

6. Click "**OK**." This will create an empty annotation that does not color any nodes. To color nodes, you must add them to the annotation as described below.

To add nodes to an annotation manually:

NOTE: This operation can only be performed if an annotation has already been created.

1. Click the "**Add Nodes**" button in the **Annotations** task panel

2. Select the desired annotation by clicking on it in the list box.

3. Click on nodes in the tree panel.

   If a selected node is a leaf node, it will be highlighted with the color of the annotation. If it is an internal node, all of leaf nodes below it will be highlighted with the color of the annotation.

4. If you want to add nodes to another annotation, repeat steps 2 and 3.

5. When you are finished adding nodes to annotations, click the "**Add Nodes**" button again to deselect it.

To remove nodes from an annotation manually:

NOTE: This operation can only be performed if an annotation has already been created and nodes have been assigned to it.

1. Click the "**Remove Nodes**" button in the **Annotations** task panel, and then select the desired annotation from the list of annotations.

2. Click on nodes in the tree panel.

> If a selected node is a leaf node, it will be removed (*i.e.*, disassociated) from the annotation and the color of its label will revert to black or the color of another annotation with which they are associated. Clicking on an internal node removes all of the leaf nodes in the subtree rooted at that node.

3. When you are finished, click the "**Remove Nodes**" button again to deselect it.

To edit an annotation:

1. Select the annotation you want to edit in the list box.
2. Click the "**Edit**" button in the task panel. The annotation dialog box will reappear. You can now change the color, title or pattern matching term(s) for this annotation.
3. Click "**OK**" when you are done making changes.

To hide/view an annotation:

1. Select the annotation in the list box.
2. Click the "**Show/Hide**" button.

> If the annotation was displayed prior to clicking the "**Show/Hide**" button, the nodes associated with the annotation will revert to black or the color of another annotation with which they are associated. If the annotation was hidden, the associated nodes will appear in color. A check mark next to the annotation's name denotes that it is visible (*i.e.*, the current state is "Show"). This is the default status.

To delete an annotation:

This function will remove an annotation from the list of annotations. All nodes associated with it will revert to black or the color of another annotation with which they are associated. *Warning: this operation is not reversible.*

1. Select the annotation in the list box.
2. Click the "**Delete**" button in the task panel.
3. A dialog box will appear to confirm that you really want to delete the annotation. Click "**Yes**."

To export an annotation:

Annotations can be exported to a separate file for import into another tree.

1. Click "**File → Export Annotations**."
2. A file dialog box will appear. Enter a name for the annotation, and click "**Save**." The annotations can now be imported from this file to another tree.

To import an annotation:

Annotations can be imported from any file that contains an annotation, including a Notung format tree or an exported annotation file.

1. Click "**File → Import Annotations.**"
2. An open file dialog box will appear, select the annotation file or annotated tree, and click "**Open**." The annotations from the file will be added to the open tree. If the open tree was previously annotated, those annotations will still be present.

   NOTE: Imported annotations are added to the existing list of annotations. If an imported annotation and a previously existing annotation correspond to the same node, the imported annotation will take precedence.

Chapter 11

Changing the Appearance of the Tree Panel

Notung offers the user a broad range of options for controlling the appearance of the tree panel and the types of information that can be displayed. Visual presentation of trees in Notung can be changed in two ways. One set of options is found in the **Display Options, Zoom**, and **Font** menus in the upper left hand corner of the Notung window. These options, which are described in this section, are relevant in all task modes.

In addition, certain visual features can be controlled from individual Task Panels. These features are typically specific to that task and, in most cases, are only visible when the relevant Task Panel is selected. These options, such as highlighting swappable nodes or weak edges and displaying root scores, are described in the relevant mode sections.

11.1 Display Options

Notung allows users to show or hide node and edge labels using the Display Options menu. Checkboxes next to each item in the menu show which display options are turned on. These options are tree-specific: changing them in the currently selected tree will not change them in other open trees.

To turn on/off a display option:

- Click Display Options, then click the desired option.

√ **Display Edge Weights (default : ON)** When this option is turned on, the weight of each edge is displayed in green type below the edge.

√ **Display Loss Nodes (default: ON)** When this option is turned on, the implied lost nodes are displayed in the tree.

√ **Display Duplications (default: ON)** When this option is turned on, a red D appears next to each duplication node.

√ **Display Co-Divergences (default: ON)** When this option is turned on, a pink cD appears next to each co-divergence.

√ **Display Transfers (default: ON)** When this option is turned on, a yellow T appears next to each transfer edge.

√ **Display Leaf Node Names (default: ON)** When this option is turned on, leaf node names are displayed.

√ **Display Internal Node Names (default: OFF)** When this option is turned on, the name of each internal node appears in red type to the upper left of the node.

> *NOTE*: If internal nodes are not labeled in the input gene tree file, Notung assigns internal node names using a counting system. Node names derived this way begin with n or r, followed by number signifying the order in which the node was counted (*e.g.*, $n136$ or $r122$). Node names beginning with an r indicate that the node did not exist in the original tree; rather, they were added during rearrangement of weak edges or resolution of a polytomy. These node names may change if other tasks are performed after the rearrangement or resolve task that produced the new internal node(s). Counting may be different for each session, so node names may vary from session to session, depending on how many trees have been opened in Notung.

√ **Display Leaf Node Species Names (default: OFF)** When turned on, this option displays the name of the species associated with each leaf node. It appears in black italic text next to the node.

> *NOTE:* In previous versions of Notung, punctuation in species names was used to indicate that Notung should search for a shorter species tag in gene names, rather than searching for the entire species name. This functionality has been removed in Notung 2.6 and beyond. For more information, see Appendix A.5 - Punctuation in Species Names on page 130.

√ **Display Internal Node Species Names (default: OFF)** When this option is turned on, the (ancestral) species associated with each internal node in the tree is displayed. It appears in black italic text to the right of the node.

> *NOTE:* If ancestral nodes are not labeled in the input species tree, internal nodes are assigned numerical identifiers. Since the species name is derived

from the reconciliation, if the gene tree has not been reconciled, this option has no effect.

√ **Highlight Polytomies (default: OFF)** When turned on, this option highlights and circles in cyan each vertical edge representing a polytomy.

√ **Use Species Names in Polytomy Losses (default: ON)** When reconciling a gene tree with a non-binary species tree, if orthologous genes are absent in two or more children of a polytomy in the species tree, a single loss in an ancestral species is inferred (called a polytomy loss). When this option is turned on, polytomy losses are displayed with the names of the species in which the gene is absent, as well as name of the polytomy (see Figure 5.3(a)). When turned off, combined losses are labeled with the number of species in which the gene is absent, and the name and the size of the polytomy (Figure 5.3(b).)

11.2 Zoom

Notung allows users to zoom in on the tree using either the Zoom menu or keypad controls. Users can zoom in on the whole tree, maintaining the tree's aspect ratio, or on the X or Y axis independently, elongating the vertical or horizontal edges, respectively. These changes apply only to the currently selected tree.

To zoom in on the whole tree:

- Click "**Zoom → Zoom In**."
  -or-
- Press "**Ctrl/Cmd**" and click on the region of interest in the tree.

  NOTE: Ctrl/Cmd means to use the "Control" key for Windows or Linux, and the "Command" (open apple) key for Mac. For operations which do not involve clicking on the tree, the "Control" key is used for all three platforms.

To zoom out on the whole tree:

- Click "**Zoom → Zoom Out**."
  -or-
- Press "**Shift**" and click on the tree.

To zoom in on the X axis:

- Click "**Zoom → Zoom In on X axis**."
  -or-
- Press "**Ctrl -**]"

To zoom out on the X axis

- Click "**Zoom → Zoom Out on X axis**."
  -or-
- Press "**Ctrl -** ["

To zoom in on the Y axis:

- Click "**Zoom → Zoom In on Y axis**."
  -or-
- Press "**Ctrl - Shift -**]"

To zoom out on the Y axis:

- Click "**Zoom → Zoom Out on Y axis**."
  -or-
- Press "**Ctrl - Shift -** ["

To fit the whole tree in the tree panel:

- Click "**Zoom → Show Whole Tree**."
  -or-
- Press "**Ctrl - T**"

11.3 Changing Font Size

Users can modify the font size of tree labels using the Fonts menu or keypad controls. Fonts can be set to one of four sizes or changed incrementally.

To set a font size:

- Click "**Fonts**," then click the desired size.
  - Tiny fonts
  - Small fonts
  - Medium fonts (*default*)
  - Large fonts

To increase font size incrementally:

- Click "**Fonts → Increase Font Size.**"
  -or-
- Press "**Ctrl - =**"

To decrease font size incrementally:

- Click "**Fonts → Decrease Font Size.**"
  -or-
- Press "**Ctrl - Minus**"

NOTE: Selecting "**Large fonts**" does not display the largest possible font; the font can be made even larger by using the "**Increase font size**" option.

Chapter 12

Command Line Options and Batch Processing

Notung offers a command line interface that can perform most operations from the command line without launching the graphical user interface. The command line interface allows the use of batch processing to apply Notung to many trees in a large-scale analysis without human intervention. It can also be used to analyze a small number of trees without launching the GUI, for example, by a user executing Notung on a remote computer over the network. The GUI can also be launched from the command line, rather than by clicking on an icon, allowing the user to initiate the GUI with parameter settings other than the default settings.

We follow the following stylistic conventions in this chapter.

- Text that should be typed verbatim on the command line, such as the name of the Notung executable and program options, is given in `teletype` font; (e.g. `Notung-2.9.jar --reconcile ...`.)
- Throughout this chapter, we refer to the Notung executable using the most recent major version number (i.e., `Notung-2.9.jar`). We periodically release versions of Notung with minor changes, such as bug fixes. The minor version number is increased with each new minor release. You may have a more recent build with a minor version number, (e.g., `Notung-2.9.1.jar`.) If so, when typing the commands given in the examples in this chapter, you should replace `Notung-2.9.jar` with name of the Notung executable in the distribution that is currently installed on your computer.
- Text in angle brackets indicates that the user must supply a parameter value. For example, `Notung-2.9.jar --reconcile -g <genetree>`, indicates that Notung expects a file name after `-g`.

- Angle brackets are also used to describe fields that will be instantiated with a value during Notung's execution. For example, many output files include either `reconcile, rearrange, resolve,` or `root` in their name, to indicate the task used to analyze the original gene tree. Reconciling the gene tree stored in a file called `mygenetreefile` will produce the file `mygenetreefile.reconciled.ntg`. In this chapter, we use `<function>` to describe such file names, e.g., `<genetree>.<function>.ntg`.
- When Notung outputs more than one optimal tree, it generates multiple output files, numbered starting at zero, yielding file names such as `mygenetreefile.reconciled.0.ntg`, `mygenetreefile.reconciled.1.ntg`, etc. We use a '`#`' sign to represent the number in such file names; e.g., `<genetree>.<function>.#.ntg`.

12.1 Opening and Using a Command Window/Terminal

Prior to running Notung's command line interface, you will need to open a command or "terminal" window.

On Windows XP

Opening a command window

Click on the Start button, and select the "Run..." item. A dialog box will pop up. Enter "`cmd.exe`" into the box, and click "OK."

Navigating to the Notung directory

In the command window, type the following

```
cd   <pathname>
```

where `<pathname>` is the path of the Notung directory. If the folder location has any spaces in it, it must be enclosed in quotes. For example, if the following is the location of the Notung folder:

```
C:\Documents and Settings\User\Desktop\Notung-2.9
```

Then you should use quotes so that it looks like this in the command window:

```
cd "C:\Documents and Settings\User\Desktop\Notung-2.9"
```

Hit Enter, and you will now be in the Notung Folder.

> *NOTE*: To find the path of the Notung directory, select the Notung folder in Explorer, and right click on it. This will pop up a menu - select the Properties item. This will pop up a dialog listing the properties of the Notung folder, including its location.

On Windows Vista or later

Opening a command window in the Notung directory

Select the Notung folder in Explorer, and SHIFT + right-click on it. This will pop up a menu - select "Open command window here."

On Mac

Opening a terminal

The Terminal application is located in the Applications folder in the Utilities subfolder.

Navigating to the Notung directory

In the terminal window, type the following

```
cd <pathname>
```

where `<pathname>` is the path of the Notung directory. If the folder location has any spaces in it, it must be enclosed in quotes. For example, if the following is the location of the Notung folder

```
/Users/user/Desktop/New Folder/Notung-2.9
```

Then it should look like this in the terminal window

```
cd "/Users/user/Desktop/New Folder/Notung-2.9"
```

Hit Enter, and you will now be in the Notung Folder.

> *NOTE*: To find the path of the Notung directory, select the Notung folder in the Finder, and select "Get Info" from the File menu. This will pop up a dialog listing the properties of the Notung folder, including its location. You could also drag and drop the Notung folder into the Terminal window to paste the folder's path into the window.

On Linux

Navigating to the Notung directory

In the terminal window, type the following

```
cd  <pathname>
```

where `<pathname>` is the path of the Notung directory. If the folder location has any spaces in it, it must be enclosed in quotes. For example, if the following is the location of the Notung folder

```
/Users/user/Desktop/New Folder/Notung-2.9
```

Then it should look like this in the terminal window

```
cd "/Users/user/Desktop/New Folder/Notung-2.9"
```

Hit Enter, and you will now be in the Notung Folder.

12.2 Running Notung from the command line

Notung can carry out its four main tasks: reconcile, rearrange, root and resolve, from the command line. In each case, Notung reads in gene and species trees (the input trees) and executes the specified task, resulting in one or more modified trees (the output tree(s)). This modified tree is written to a file. Notung can also generate images in PNG format from the command line. This function can be carried out in conjunction with any of the four main tasks, or independently to generate an image of an existing tree without performing any analysis. The I/O requirements differ somewhat in the latter case; only one tree is required as input and an image rather than a tree file is generated as output. In this section, we discuss executing the four main tasks from the command line. Commands and options specific to image generation are described in Section 12.5. In Section 12.3, automated execution of Notung is described. Notung can gather useful phylogenomic analysis data from batch reconciliations. These statistics are described in Section 12.4. Commands and options specific to reconciliation with non-binary species trees are described in Section 12.6.

For the four major tasks, Notung is executed from the command line using the following format:

```
java -jar Notung-2.9.jar [input tree(s)] [task] [options]
```

The four main tasks require both a gene tree and a species tree. These are usually supplied as two separate input files. A single file containing a previously reconciled tree in Notung format is also acceptable, since such files contain both a gene tree and species tree (see Appendix A.3 - Notung File Format on page 128). If a gene tree file containing a reconciled tree in Notung format *and* a species tree in a separate file are both given, the latter is used; the species tree in the gene tree file is ignored. The task parameter must be one of `--reconcile`, `--rearrange`, `--root`, or `--resolve` (the fifth task, `--savepng`, is discussed in Section 12.5 - Saving PNG Images of Trees.) Options are described at the end of this chapter.

NOTE:

- The input trees, tasks, and options may be given in any order.
- To launch the graphical interface from the command line, run Notung without a task option and without a command to save an image.

  ```
  java -jar Notung-2.9.jar
  ```

- Running Notung with the `--help` option causes it to print information regarding input, output, and other options.

- You may also run Notung anywhere from the command line by giving the path during launch:

```
java -jar <pathname>/Notung-2.9.jar [input tree(s)] [task] [options]
```

  Alternatively, you can add the Notung directory to your CLASSPATH. For example, if you run bash in Linux, you can do this by adding the following command to your .bashrc file:

```
setenv CLASSPATH $CLASSPATH:<pathname>
```

  See a java manual for more information about CLASSPATH settings.

The following list summarizes the types of information that Notung will generate, with a brief description of each output file type. File formats for the information generated with the `--phylogenomics` option are described in Section 12.4 - Running a phylogenomic analysis. For more details on tree formats, including information on edge weights, species tags, and output files, see Appendix A - File Formats on page 126.

Output options

Output Gene Tree(s) If one of the four main functions is given, the output gene tree will be saved to a file called `<genetree>.<function>.<extension>`, where `<function>` is one of the four major tasks and `<extension>` is `ntg` for Notung format, `nhx` for Newick Extended format, and `nwk` is Newick or New Hampshire format. If the analysis results in more than one optimal history, then the output files are numbered, (*e.g.* `<genetree>.rearrange.0.<extension>`, `<genetree>.rearrange.1.<extension>`, *etc.*). By default only one tree is saved. To save more than one history when inferring transfers, use `--multsols`. To save more than one history for all other situations (rearranging, rooting, or non-binary gene trees), use `--maxtrees`.

> *NOTE:* For the remainder of the manual, we will use `.ntg`, the default, as the tree file extension, but `.nwk` and `.nhx` are other options. To change the tree format, use `--treeoutput`.

PNG Image of Tree (optional) If the `--savepng` or `--saveweakedgespng` option is given, an image of the tree is saved in PNG format. For more information on saving PNG images, see Section 12.5 - Saving PNG Images of Trees on page 111.

Pruned Species Tree (optional) In some cases, Notung prunes species that are not represented in the gene tree from the species tree prior to reconciliation. There are two possible pruning strategies, described in detail in Section 5.1. When the `--stpruned` option is given, the pruned species tree is saved in the file `<genetree>.<function>.species.ntg`.

Pruned Gene Tree (optional) If the gene tree contains species taxa not represented in the associated species tree, Notung will print an error message and terminate. If you wish to ignore these taxa in the gene tree, you can use the `--gtpruned` option on the command line. This will prune the gene tree to remove all taxa without a representative in the species tree. Use caution when using this option, as taxa could be pruned due to incorrect labeling rather than a true missing species.

Log (optional) When run on the command line, Notung outputs status information to the terminal window. This information can be saved in a log file `<genetree>.<function>.ntglog` by using the `--log` option. For a batch run, a log file is not saved for each tree; rather, a single log file for the entire batch run is saved to the file `<batchfile>.<function>.ntglog`.

General Tree Statistics (optional) General tree statistics can be saved in the file `<genetree>.<function>.stats.txt` by giving the option `--treestats`. This file includes information on both the gene tree and the associated species tree. For more information on tree statistics, see Section 3.4 - General Tree Statistics on page 20.

Event Summary (optional) Information on each duplication, transfer, co-divergence, and loss is saved in the file `<genetree>.<function>.events.txt` when the `--events` option is used. For each duplication and co-divergence, upper and lower bounds on the time of the event (represented as nodes from the species tree) are given. For transfers, the donor and recipient species (represented as nodes in the species tree) are supplied. For losses, each node in the species tree is listed with the number of losses associated with that taxon. At the end of the file, a table is provided, where each row represents a node in the species tree and each number in a row indicates the number of times a duplication, co-divergence, or loss occurred in that species, as well as the number of times that species acted as a donor or recipient of a transfer. For more information on the event summary, see Chapter 5.5.1 - Event Summary on page 51.

Parsable Statistics (optional) When the `--parsable` option is used, information on inferred events is saved to an easily parsable, tab-delimited file `<genetree>.<function>.parsable.txt`.

To facilitate automated postprocessing by scripts, lines relating to bounds for duplications and co-divergences are preceded by #D and #CD, respectively; while lines relating to transfer donors and recipients are preceded by #T. Each line of the loss table, as in **Event Summary**, is preceded by a #L. The summary count table over all species and events is denoted by #S. For more information on the parsable statistics file, see Chapter 5.5.1 - Parsable Statistics on page 52.

Homology Tables (optional) Notung can output tables of orthologs, paralogs, and xenologs for all pairs of leaf nodes in the reconciled tree. This table can be generated in several formats: comma-separated values (CSV), tab-delimited values, or an html-formatted table. Use options `--homologtablecsv`, `--homologtabletabs` or `--homologtablehtml`, respectively. For more information on orthologs, paralogs, and xenologs, see Section 5.6 - Inferring Orthologs, Paralogs, and Xenologs on page 53.

12.3 Running Notung from a Batch File

Batch processing allows the user to apply Notung to many trees in a large-scale, automated analysis. The input trees are given in a batch file, which consists of a list of tree file names, one per line. Blank lines and lines which start with `#` are ignored.

To create a batch file:

- The first line of the file should give the path of the species tree to be used.
- Each subsequent line should give the path of a gene tree.

> *NOTE:* By default, Notung expects the tree file locations to be given relative to the location of the batch file. For example, if the batch file is in `/username/batchRun`, Notung expects the gene trees to be in the `batchRun` folder or in some subfolder of `batchRun`. Use the `--absfilenames` option to indicate that file names are absolute path names.
>
> *NOTE:* When using the `--savepng` option without any of the four main functions (`--reconcile`, `--root`, `--rearrange` and `--resolve`), each tree listed in the batch file is saved as an image. For more information, see Section 12.5 - Saving PNG Images of Trees on page 111.

A sample batch file is provided with the Notung-2.9 distribution in the `sampleTrees/batch` directory. This batch file includes all combinations of binary and non-binary gene and species

trees. Because not all of Notung's task modes work for each of these combinations, you will receive one or more warnings and errors when running this batch file. In addition, the batch file lists a gene tree which does not exist, to give an example of the appropriate warning.

To run Notung from a batch file:

Use the `-b <batchfile>` option.

For example, from the Notung directory, enter the following on the command line:
`java -jar Notung-2.9.jar -b sampleTrees/batch/batch.run --reconcile --speciestag prefix`
The `--reconcile` option tells Notung to reconcile all the gene trees listed in `batch.run` with the species tree listed in `batch.run`. The `--speciestag` prefix option tells Notung how species labels are specified in the gene tree files, and is required in batch mode. See Appendix A.4 - Specifying the Species Associated with Each Gene on page 129 for more information on species labels.

> *NOTE*: All gene trees in the same batch file must use the same species tag format, which is specified using the `--speciestag` option.

Required Options

In batch mode, the `--speciestag` option is always required. In addition, when using `--rearrange`, `--edgeweights` and `--threshold` must be used to set the edge weight locations and threshold, respectively.

Batch Output

As Notung reads and processes each gene tree in the batch file, it prints diagnostic information to the terminal. Notung will also print this information to a log file when the `--log` option is given. Any errors that occur in the processing of a batch file are reported to the terminal as they occur. The total number of errors is reported at the end of the batch run.

To print status information to a file:

Use the `--log` option from the command line. The information will then be written to the file `<batch_file_name>.ntglog`.

To save trees to a different directory:

By default, Notung saves each reconciled tree to the directory from which the program was run.

- To save trees to a different directory, use the `--outputdir <outputDir>` option from the command line. The information will then be written to the directory `<outputDir>`.
- Alternately, the `--usegenedir` option can be used to save output to the directory in which each gene tree is located.

Progress Bar

For long runs, it may be convenient to use the options `--silent` and `--progressbar` together. This will suppress all output to the terminal with the exception of a simple progress bar to stderr. The option `--log` can still be used to save the (now suppressed) output to a file.

12.4 Running a phylogenomic analysis

The evolutionary history of a genome can be considered from the perspective of the combined evolutionary histories of its constituent gene families. With the growing availability of whole genome sequence, it is now possible to construct trees for a comprehensive set of gene families drawn from a given set of genomes. Notung-2.9's `--phylogenomics` option aggregates the events and ancestral states inferred for individual trees to reveal genome scale trends, such as

- an enrichment of duplications associated with a given ancestral species, suggestive of a whole genome duplication;
- a wave of losses characteristic of genome reduction;
- co-evolution of gene families with related functions;
- bursts of horizontal gene transfer between particular pairs of species;
- stratigraphic analysis of gene age relative to the species tree.

Wagner parsimony, applied to the number of genes (or gene families) found in present day genomes, is a common approach to asking such questions. However, Wagner parsimony will

underestimate events and ancestral gene content in families that have sustained parallel gains or losses. Estimates obtained using reconciliation are more accurate because the inference is guided by the topologies of the gene trees.

Given a batch file listing a species tree and a set of gene trees, Notung-2.9's phylogenomic function reconciles each gene tree and outputs tables summarizing ancestral gene content and the combined set of duplications, losses, and transfers associated with each node or branch in the species tree. The following command is the simplest example of a phylogenomics analysis:

```
java -jar Notung-2.9.jar -b batch.txt --reconcile --speciestag prefix
--phylogenomics
```

This command will reconcile each gene tree in the batch file and save the resulting reconciled tree, as described in Section 12.2 - Running Notung from the command line above. In addition, summary statistics aggregated over all reconciliations will be collected and output in tab-delimited tables described below.

As with all analyses in `--reconcile` mode, the `--speciestag` must be used to specify the location of the species name in the gene tree leaf labels. Additional command line options can be used to change default parameter settings (e.g., the event costs) or to generate additional output files (e.g., `--events` or `--parsable`).

Phylogenomic analysis with transfers requires the `--infertransfers true` option. The phylogenomics function is only defined for batch analysis in `--reconcile` mode, executed from the command line. It is not defined for `--rearrange`, `--root`, or `--resolve` mode. A Notung-2.9 command that pairs one of these modes with `--phylogenomics` will generate an error.

Phylogenomics output

The `--phylogenomics` command aggregates event statistics over all reconciliations and outputs tab-delimited tables with summary statistics. With the exception of the `<batchfile>.highway.txt` file, described below, all tables have the same format: one row for each gene tree in the batch file and one column for each node (including leaf nodes) in the species tree. If reconciliation is performed with the `--infertransfers` option, only trees that have at least one temporally feasible solution (see Chapter 5.2.1 - Temporal Feasibility in Notung on page 43) are included in the table.

Gene family origins:

The `<batchfile>.<speciestree>.origin.txt` file reports the origin of each gene family; that is, the species tree branch in which the family was first observed. Stated technically, the origin is defined to be the species node associated with the root of the reconciled gene tree. The value of entry i, j in this table is 1 if family i arose in species j and 0 otherwise.

Ancestral gene content:

The `<batchfile>.<speciestree>.geneCount.txt` table reports the inferred ancestral gene content in each species tree node. Entry i, j is a non-negative integer representing the number of members of gene family i present in species j. The total gene content of each species corresponds to the column sums of this table. The number of gene families that were present in each species can be determined by counting the number of non-zero entries in each column.

Gain and loss of gene families:

The `<batchfile>.<speciestree>.famGainLoss.txt` file reports for each gene family, all branches in the species tree where the family was gained or lost. The value of entry i, j in this table is 1, if family i was present in species j, but absent in the parent of j; -1, if family i was present in the parent of j, but absent in j; and 0 otherwise. This table includes the gain at the origin of the gene family; that is, if entry i, j is 1 in `<batchfile>.<speciestree>.origin.txt`, it will also be 1 in `<batchfile>.<speciestree>.famGainLoss.txt`. To obtain only those cases where the family was lost and then regained, subtract the entries in `<batchfile>.<speciestree>.origin.txt` from the entries in `<batchfile>.<speciestree>.famGainLoss.txt`.

Net change in gene family size:

The `<batchfile>.<speciestree>.geneGainLoss.txt` file reports the change in gene copy number in each family for every branch in the species tree. The value of entry i, j in this table is $n > 0$, if the number of members of gene family i in species j has increased by n, compared with the number of members in the parent of species j. If n is a negative number, then the number of members of gene family i in species j has decreased by n, relative to the count in the parent of j. If entry i, j is zero, then there was no net change.

Duplications:

The `<batchfile>.<speciestree>.duplication.txt` table reports the number of inferred duplications in each species tree branch. Entry i, j is a non-negative integer representing the number of duplications that occurred on the branch from the parent of species j to j.

Losses:

The `<batchfile>.<speciestree>.loss.txt` table reports the number of inferred losses in

each species tree branch. Entry i, j is a non-negative integer representing the number of losses that occurred in family i on the branch from the parent of species j to j.

Transfers:

The `<batchfile>.<speciestree>.transferFrom.txt` table summarizes outgoing transfers. Entry i, j is a non-negative integer representing the number of transfers in family i that originated in species j, summed over all possible recipients.

The `<batchfile>.<speciestree>.transferTo.txt` table summarizes incoming transfers. Entry i, j is a non-negative integer representing the total number of transfers in family i that originated from any species and were received by species j.

These tables are only generated when `--infertransfers` is set to `true`.

Transfer highways:

The final table generated with the `--phylogenomics` option has one row for each species and one column for each species. The `<batchfile>.<speciestree>.highway.txt` table summarizes the total number of transfers, summed over all gene families, between pairs of species. Entry i, j is a non-negative integer representing the number of transfers with donor species i and recipient species j. This table is only generated with the `--infertransfers true` option.

Phylogenomics with averaging

Recall that with transfers, a gene tree may have zero, one, or more than one optimal, temporally feasible reconciliation (see Chapter 4.1.1 - Temporal Feasibility on page 28 for more details). Only gene trees that have at least one temporally feasible reconciliation are included in the tables generated by the `--phylogenomics` option. If a gene tree has more than one optimal, temporally feasible reconciliation, one of those reconciliations is selected arbitrarily for tabulation in the summary tables. The averaging feature allows the user to extract information from all optimal, temporally feasible reconciliations.

This command exemplifies the minimal set of options required for a phylogenomics analysis with averaging:

```
  java -jar Notung-2.9.jar -b batch.txt --reconcile --speciestag prefix
--phylogenomics --avg --infertransfers true
```

The `--avg` option can only be used with batch analysis in `--reconcile` mode with the `--phylogenomics` and `--infertransfers true` options. Averaging is only defined when transfers are included in the event model; with duplications and losses, alone, the optimal reconciliation is unique.

This command generates the tables described above in Phylogenomics output, but information from all optimal reconciliations of each gene tree is used to calculate the summary statistics in those tables. Specifically, for a given gene tree, the values obtained by averaging over all feasible, optimal reconciliations obtained with that gene tree. Thus, with `--phylogenomics --avg`, the entries in these tables are real numbers, in contrast to analyses with `--phylogenomics` alone, which generate tables with integer values.

In each table, except the **Transfer highways** table, entry i, j is the sum of the values obtained with all optimal, temporally feasible reconciliations of gene family i, normalized by the total number of such reconciliations. In `<batchfile>.<speciestree>.highway.txt`, entry i, j is the average, over all gene families, of the mean number of transfers from species i to species j, where the mean number of transfers experienced by a gene family is obtained by averaging over all optimal reconciliations of the gene tree for that family. In other words, although some gene trees may have more optimal reconciliations than others, each gene family is weighted equally. As the number of optimal reconciliations increases, the contribution of each reconciliation decreases accordingly.

Note that the command

```
  java -jar Notung-2.9.jar -b batch.txt --reconcile --speciestag prefix
--phylogenomics --avg --infertransfers true
```

saves (at most) one reconciled tree for each gene tree in `batch.txt`, but enumerates all optimal solutions during the averaging step. This will increasing the running time of the analysis.

12.5 Saving PNG Images of Trees

The option `--savepng` saves a simple image representation of a tree in PNG format, while the option `--saveweakedgespng` saves a simple PNG representation of a tree with weak edges highlighted in yellow. A threshold defining weak edges must be given using `--threshold` with the latter option; if not, a default of 90% will be used. These options can be used with one of the four main tasks (`--reconcile`, `--root`, `--rearrange` and `--resolve`), in which case an image of the final output tree is saved, in addition to the output tree file. To save

an image of the tree without losses, use the `--nolosses` command. Alternatively, they can be used alone to save an image of a tree without performing any other tasks.

Using `--savepng` and `--saveweakedgespng` alone

When these options are used without one of the main four tasks, Notung reads in a tree and generates and saves an image of that tree in PNG format. Unless a batch file is used, only a single tree can be processed at a time (*i.e.*, a gene tree and a species tree *cannot* both be given). If the input tree is a previously reconciled tree in Notung format, the image will show the appropriate events (to save an image without losses, use `--nolosses`). If the tree has not been reconciled, the tree image will show only the structure of the tree and the names of the leaves of the tree.

When using a batch file, each tree specified in the file is saved as an image. When generating images without performing a major task, the batch file format differs slightly: Species trees and gene trees can be listed in any order.

Output File Names

When `--savepng` or `--saveweakedgespng` is used alone, an image of the input tree is saved in the file `<treename>.png`. When used with `--reconcile`, `--root`, `--rearrange` or `--resolve`, an image of the output tree is saved in the file `<genetreename>.<function>.png`. For analyses with more than one optimal history, an image file is saved for each history. The number of files is limited by the parameters `--maxtrees` and `--multsols`.

Color Annotations

If a tree in Notung format contains color annotations, the leaves in the image of that tree will be colored as specified by those annotations. An annotation file can be specified with the option `--annotationfile`. For more information on color annotations, see Chapter 10 - Annotations on page 87.

Making an Imagemap

Notung provides the option to produce an html imagemap for a tree image. If an imagemap and image file are both included in a web page, each gene in the image will provide a link to a specified web page. The format of these links is determined by the imagemap specification

file given with `--imagemapfile <imagemapfilename>`, described below. The resulting imagemap is saved in the file `<outputtreename>.png.html`, where `<outputtreename>` is either `<genetree>.<function>` or `<treename>`.

To include the image and imagemap in a web page, insert the entire contents of the saved imagemap file into the html of the web page. The saved image must be in the same directory as the web page, unless you specify a different location for the image by changing `<imagefile>` in the line:

```
<img border=0 src='<imagefile>' ...
```

Imagemap Specification

The specification file given by `--imagemapfile <imagemapfilename>` consists of a list of gene-link pairs. Blank lines and lines that start with `#` are ignored. Entries creating the imagemap link follow the format:

```
# Descriptive comment:
gene: <search string>(id)
link: <link start>(id)<link end>
```

The `<search string>` exactly matches strings in the gene tree. When a match is found, the identifier `(id)` will capture all remaining text following the matched string. That text can be used in the link, specified in the '`link:`' line, thus creating a custom link for that gene.

For each gene in the gene tree, the first gene-link pair that matches will be used. If a gene does not match any of the '`gene:`' lines, a warning will be printed. To prevent this warning, the user can add the following to the end of the imagemap specification file:

```
# generic imagemap - everything else links to google
gene: (id)
link: http://www.google.com/search?q=(id)
```

Any leaf in the gene tree that does not match a previous entry will match this entry, and will link to a Google search for that leaf string.

An example gene tree and imagemap specification from the Princeton Protein Orthology Database (ppod.princeton.edu/) are included in the Notung distribution. An sample specification file for this example follows:

```
# Danio rerio links:
gene: Danio_rerio|(id)
link: http://zfin.org/cgi-bin/ZFIN_jump?record=(id)

# generic imagemap - everything else links to google
gene: (id)
link: http://www.google.com/search?q=(id)
```

The gene `Danio_rerio|ZDB-GENE-031007-1` would match the first '`gene:`' line, with `(id)` equal to `ZDB-GENE-031007-1` and with link `http://zfin.org/cgi-bin/ZFIN_jump?record=ZDB-GENE-031007-1`. The gene `Homo_sapiens|gene1` would match the second pair, because '`(id)`' will match any text string. The resulting link would be `http://www.google.com/search?q=Homo_sapiens|gene1`.

12.6 Inferring Losses when Reconciling with Non-Binary Species Trees

When inferring losses during reconciliation with a non-binary species tree, it is not possible to determine unambiguously the edge in the the gene tree to which a loss should be assigned. Notung uses two different methods to deal with this problem. An exact algorithm finds all possible assignments that minimize the total number of losses but has exponential time complexity. A heuristic, which runs in polynomial time, is not guaranteed to find the optimal assignment, but usually does in practice. These issues and algorithms are discussed in detail in Section 4.2 - Non-Binary Trees.

Only the heuristic is implemented in the GUI and for the transfer algorithm. Either method may be used when executing Notung from the command line without the `--infertransfers` option. The CLI runs the heuristic by default. To use the exact algorithm, include the `--exact-losses` option when running Notung from the command line with the `--reconcile` or `--root` tasks.

The running time of the exact algorithm is exponential in the size of the largest polytomy. Even when `--exact-losses` is used, Notung does not apply the exact algorithm to polytomies with more than 12 children. Instead, the heuristic is applied to these polytomies. To change the maximum polytomy size for which Notung uses the exact algorithm, use the `--polytomy-cutoff <maxPolytomySize>` option when including the `--exact-losses` option in the command line.

NOTE: Changing the polytomy cut-off to a larger value and using the exact

algorithm on a species tree with a polytomy with more than 12 children may greatly increase running time.

Command Line Options for Losses with Non-Binary Species Trees

`--exact-losses`

Computes the minimum number of losses when reconciling a binary gene tree with a non-binary species tree. If this option is not included on the command line, the heuristic is used. `NOTE:` In Notung 2.5, this option was named `--combine-losses`.

`--polytomy-cutoff <maxPolytomySize>`

Using this option with `--exact-losses` will change the default value for polytomy cutoff. Only for losses associated with polytomies less than or equal to `<maxPolytomySize>` will the exact algorithm be used. The default value is 12. If a polytomy greater than `<maxPolytomySize>` is encountered, a warning will be printed to the terminal window and/or log file.

`--report-heuristic-losses`

When run with `--exact-losses`, this option will report both the number of losses obtained with the heuristic and with the exact algorithm. This is useful for determining whether the heuristic is overestimating the number of losses and by how much. `NOTE:` In Notung 2.5, this option was named `--report-explicit-losses`.

12.7 Rooting trees from the command line with the DTL model

In Root mode, by default Notung finds the set of edges with optimal a root scores, when `--infertransfers` is set to "true." This stands in contrast to the rooting analysis with the DL model, where the root score is calculated for every edge in the gene tree. With the DTL model, the analysis is restricted to edges with optimal root scores because the transfer model incurs the relatively high computational complexity of inferring candidate reconciliations and testing temporal feasibility.

There may be more than one edge with the minimum root score and there may be more than one optimal event history for each such edge. Under default settings, the command line version outputs the minimum root score and saves a reconciled tree that is rooted on one of the optimal edges, chosen arbitrarily. Various command line options allow the user

to save additional optimal rooted, reconciled trees. To see additional optimal roots, use the `--maxtrees` and `--allopt` flags. The options

```
--maxtrees <n> --allopt
```

will save trees rooted on all edges with the optimal root score, or on `<n>` such edges, if there are more than `<n>` optimal roots. If the `--maxtrees` flag is used without `--allopt`

```
--maxtrees <n>
```

then Notung will save `<n>` trees rooted on the `<n>` highest scoring edges with feasible reconciliations. If the number of optimal roots is less than `<n>`, then trees rooted on edges with suboptimal scores will be saved.

By default, the rooted trees saved during rooting analysis are reconciled trees in Notung format. This means that they may be annotated with duplication and transfer events and will have loss nodes. This is useful for further evaluation or visualization with Notung, but may cause problems if you plan to further analyze these rooted trees with another application. To save rooted trees without events and loss nodes, use

```
--treeoutput newick --nolosses
```

Because testing multiple candidate solutions for temporal feasibility can significantly increase running time, the rooting analysis tests the feasibility of candidate optimal reconciliations until a temporally feasible reconciliation is found. This is sufficient to determine that a tree rooted on the current edge has at least one feasible reconciliation. As a result, when the rooting analysis terminates, the optimal root score and edges with at least one minimum cost, feasible solution are known, but not the number of optimal reconciliations for these optimal roots.

To see all optimal reconciliations for each optimal root, use the

```
--maxtrees <n> --allopt
```

options with `--root` to save one rooted tree for each optimal root, as described above. Then reconcile each saved rooted tree with Notung using the

```
--reconcile --multsols <n>
```

flags to generate all optimal reconciliations for each root. In order to guarantee that you will obtain all optimal roots (respectively, all optimal reconciliations), the value of `<n>` should be greater than or equal to the number of edges in the gene tree. If your gene tree is large, this may result in very long runtimes.

12.8 Analyzing non-binary gene trees with the DTL model

Weakly-supported edges, as indicated by low edge weights, and non-binary nodes in the gene tree often imply that the inferred history associated with those edges/nodes may not be accurate. Notung can rearrange weakly-supported regions and resolve non-binary nodes in a gene tree to produce alternate event histories with minimum Event Score. When reconciling a non-binary gene tree, edges that were added to resolve the polytomy are removed before saving the reconciled tree. Notung can perform these functions using the DTL event model on the command-line. See Section 4.2.2 - Fitting a Non-Binary Gene Tree to a Binary Species Tree on page 35 for more information.

Multiple Optimal Solutions There may be more than one rearrangement or resolution with the minimal Event Score; in addition, under the DTL model, there may be more than one optimal event history for a given gene tree rearrangement or resolution. By default, Notung outputs only one minimal rearrangement/resolution and associated event history, chosen arbitrarily. As with rooting (Section 12.7), command line options allow the user to save additional optimal solutions. The option

```
--maxtrees <n>
```

will save all rearranged/resolved trees and their histories, or `<n>` trees and histories if there are more than `<n>` optimal solutions.

> *NOTE*: Currenlty, Notung does not distinguish betweeen different optimal topologies and different optimal event histories for the same topology.

To Rearrange with the DTL model Use the `--reconcile` command option with the flag `--infertransfers true`. Since support for edges is determined by edge weight, Notung's rearrangement function requires that the gene tree include edge weights, and that the `--threshold` flag be set. If there are multiple values assigned to edges, you should use the `--edgeweights` flag to set the proper edge weight locations.

To Resolve Non-binary Nodes with the DTL model Use the `--resolve` command option with the flag `--infertransfers true`.

To Reconcile a Nonibinary Gene Tree using the DTL model Use the `--reconcile` command option with the flag `--infertransfers true`.

To set the algorithm searching for the best binary representation With the DTL model, finding the best rearrangement/resolution is NP-hard. Notung provides the user with multiple options to search for the best topology.

12.9 Command line options

File Input

`-g <genetree>`

Load the file `<genetree>` as a gene tree. *NOTE*: The `-g` is optional.

`-s <speciestree>`

Load the file `<speciestree>` as a species tree. The `-s` is required.

`-b <batchfile>`

Load the trees listed in `<batchfile>`. Requires that the `--speciestag` option be set. If rearranging, requires the `--edgeweights` and `--threshold` options. With this option, `-g <genetree>` and `-s <speciestree>` should not be specified. See Section 12.3 - Running Notung from a Batch File on page 105 for more information.

`-absfilenames`

Files listed in `<batchfile>` use absolute paths. See Chapter 12.3 - Running Notung from a Batch File on page 105 for more information.

Tasks

`--reconcile`

Reconcile a gene tree with a species tree. For transfers, the top `<multsols>` best scoring histories are saved in files named `<genetree>.reconcile.#.ntg`. For non-binary gene trees, the `<maxtrees>` optimal histories are saved in files named

<genetree>.reconcile.#.ntg. By default, <multsols> and <maxtrees> are set to 1. In batch mode, --speciestag is required. For more information on reconciliation, see Chapter 5 - Reconciliation Mode on page 40.

--rearrange

Rearrange the gene tree. The top <maxtrees> best scoring rearrangements are saved in files named <genetree>.rearrange.#.ntg. By default, <maxtrees> is set to 1. The option --threshold must be set. In batch mode, --speciestag and --edgeweights are also required. For more information on rearranging gene trees, see Chapter 7 - Rearrange Mode on page 73.

--resolve

This task, which removes polytomies from a non-binary gene tree, can only be carried out if the gene tree is non-binary. The top <maxtrees> best scoring resolutions are saved in files named <genetree>.resolve.#.ntg. By default, <maxtrees> is set to 1. In batch mode, --speciestag is required. For more information on resolving non-binary nodes in a gene tree, see Chapter 8 - Resolve Mode on page 81.

--root

Root the gene tree. The top <maxtrees> best scoring rooted trees are saved in files named <genetree>.rooting.#.ntg. By default, <maxtrees> is set to 1. In batch mode, --speciestag is required. For more information on rooting gene trees, see Chapter 6 - Rooting Mode on page 68.

Reconciliation Options

If these options are not given and the gene tree has previously been reconciled, the options saved with that reconciliation will be used.

--infertransfers [true|false]

If set to true, Notung includes transfers in event inference and optimization. If set to false, Notung uses the only the duplication-loss event model. This option can only be set to true with --reconcile and --root, and only for binary gene trees. If this option is not given and the gene tree has previously been reconciled, the option saved with that reconciliation will be used. Otherwise, only duplications, co-divergences, and losses will be inferred.

--ignorelosses

When this option is given, Notung does not consider losses in the event score during event inference. Rather, events are inferred by minimizing only duplications, co-divergences, and transfers. Losses are inferred post-hoc, after these events are assigned.

`--prune`

By default, Notung does not prune species unrepresented in the gene tree, from the species tree. Use this option if you would like Notung to construct and use a species tree that does not contain species that do not appear in the gene tree. If used in batch mode, the species tree is pruned separately for each gene tree.

`--gtpruned`

This option will prune from the gene tree, all taxa with a label not found in the provided species tree. This option should be done separately, not run with other action flags (i.e., –reconcile, –root, or –rearrange).

`--costdup <duplication cost>`

Sets the cost of gene duplications. If not set, the cost is set to 1.5, by default.

`--costcodiv <co-divergence cost>`

Sets the cost of co-divergences. These only occur when reconciling a binary gene tree with a non-binary species tree. If not set, the cost is set to zero by default. See Chapter 5 - Reconciliation Mode on page 40 for more information.

`--costloss <lost gene cost>`

Sets the cost of gene losses. If not set, the default cost of 1.0 is used.

`--costtrans <transfer cost>`

Sets the cost of transfers. These are only inferred if the `--infertransfers` option is set to true. If not set, the default cost of 3.0 is used.

Input Data Options

`--speciestag [prefix|postfix|nhx]`

Indicates the format of species tags in the gene tree. If not set, Notung tries to guess the correct format. See Appendix A.4 - Specifying the Species Associated with Each Gene on page 129.

`--threshold <threshold>|<percentage>%`

Edges with weight higher than `<threshold>` are preserved during rearrangement. This can be given as an absolute value or or as a percentage of the maximum value, using `<percentage>%`; *e.g.* "`--threshold 90%`" sets the threshold at 90 percent of the highest edge weight in the tree. See Section 3.5 - Parameter Values on page 22 for more information.

`--edgeweights [name|length|nhx]`

Indicates where in the tree file the edge weights, if any, are specified. If this option is not set, and the gene tree has values in more than one location, Notung will guess the location of edge weights when using `--rearrange`. See Appendix A.6 - Location of Edge Weight Values on page 131 for more information.

`--bootstraps [name|length|nhx]`

Same setting as `--edgeweights`. Kept for backwards compatibility.

`--annotationfile <filename>`

Attach the given annotation file to each input tree.

`--imagemap <filename>`

Used with `--savepng`. Notung uses the contents of `<filename>` to create an image map file, which is saved in `<outputtreename>.png.html`. For more information, see Section 12.5 - Saving PNG Images of Trees on page 111.

Output Options

`--treeoutput [newick|notung|nhx]`

Specify output tree file format. See Appendix A - File Formats on page 126 for more information. Output trees will have the extension associated with the given format: .ntg for Notung (default), .nwk for newick, and .nhx for newick extended.

`--nolosses`

Remove loss nodes from gene trees before they are saved. Useful when outputting tree in Newick or NHX formats, which do not recognize loss nodes, or with `--savepng` to output a tree image without loss nodes.

`--maxtrees <maxtrees>`

Maximum number of optimal trees to output during reconciliation with non-binary gene trees, rearrangement, rooting, and resolving. Default is 1.

`--allopt`

In root mode, the flags `--maxtrees <n> --allopt` will find up to `<n>` optimal roots. If there are more than `<n>` optimal roots, only `<n>` rooted trees will be saved. This flag must be used with `--maxtrees`.

`--multsols <multsols>`

Maximum number of optimal trees to output during reconciliation with transfers. Default is 1.

`--outputdir <outputDir>`

Save output files in the directory, `<outputDir>`. Default is the current working directory.

`--usegenedir`

Save output trees in the directory in which `<genetree>` is located.

`--log`

Writes diagnostic output to the file `<genetree>.<function>.ntglog.txt`, where `<function>` is one of the four modes. For batch runs, the log file is saved in `<batchfile>.<function>.ntglog.txt`.

`--events`

Save information on duplications, co-divergences, transfers, and losses in the file `<genetree>.<function>.events.txt`. See Section 5.5.1 - Event Summary on page 51 for more information.

`--parsable`

Saves easily parsed information on duplications, co-divergences, transfers, losses, and general tree statistics in the file `<genetree>.<function>.parsable.txt`. See Section 5.5.1 - Parsable Statistics on page 52 for more information.

`--treestats`

Save general statistics for a tree. Saved in `<genetree>.<function>.stats.txt`. Statistics on the associated species tree will be included in this file. See Section 3.4 - General Tree Statistics on page 20 for more information.

`--stpruned`

Save a version of the species tree that contains only the species found in the gene tree. Saved in the file `<genetree>.<function>.species.ntg`.

`--rootscores`

Report a list of ordered root scores to standard output (only used with `--root`). This option is useful for statistical examination of root scores for the gene tree. These scores can be saved in a file with the `--log` option.

`--silent`

Suppresses reporting of diagnostic information to the terminal. This information may still be printed to a file if the `--log` option is used.

`--progressbar`

In batch mode, print a simple progress bar to `stderr` for each tree analyzed. Useful with `--silent`.

`--savepng`

Save the tree as a PNG image. Unlike Notung's other main functions, this function does not require a species tree. For more information about `--savepng`, see Section 12.5 - Saving PNG Images of Trees on page 111.

`--saveweakedgespng`

Saves the tree with weak edges highlighted in yellow as a PNG. This requires the `--threshold` option. If a threshold is not provided, Notung will use the default 90%. Unlike Notung's other main functions, this function does not require a species tree. For more information about `--saveweakedgespng`, see Section 12.5 - Saving PNG Images of Trees on page 111.

Homology Tables

For more information on orthologs, paralogs, and xenologs, see Section 5.6 - Inferring Orthologs, Paralogs, and Xenologs on page 53.

`--homologtablecsv`

Save a comma separated table of orthologs, paralogs, and xenologs to the file `<genetreename>.<function>.homologs.csv`.

`--homologtabletabs`

Save a tab-delimited table of orthologs, paralogs, and xenologs to the file `<genetreename>.<function>.homologs.txt`.

`--homologtablehtml`

Save a table of orthologs, paralogs, and xenologs in html format to the file `<genetreename>.<function>.homologs.html`. This format can be included in a web page.

Display Options

`--show-species-tree`

GUI only: if an input gene tree is reconciled, open the attached species tree in a separate tab.

`--homologgui`

GUI only: if an input gene tree is reconciled, start Notung in the Reconciliation tab with the Show Homology button selected.

Help Message

`--help`

Print information about these options.

Example commands

```
java -jar Notung-2.9.jar -g GENETREE.nwk -s SPECIESTREE.nwk
--reconcile --speciestag prefix --infertransfers true --costdup
 4.0 --costtrans 6.0 --costloss 2.5 --parsable
```

This command reconciles the Newick file `GENETREE.nwk` (see Section A.1 - Newick File Format on page 127 and see Section A.4 - Specifying the Species Associated with Each Gene on page 129) with the Newick file `SPECIESTREE.nwk` using custom costs and DTL reconciliation. The reconciled gene tree in Notung format (the default) will be saved in the file `GENETREE.nwk.reconciled.0.ntg`. The `--speciestag prefix` flag indicates that species names in the gene tree are given in prefix format (Section A.1 - Newick File Format and Section A.4 - Specifying the Species Associated with Each Gene). A summary of the inferred events will be saved in an easily parsable format (5.3) in the file `GENETREE.nwk.reconciled.0.parsable.txt` (See Chapter 5 - Reconciliation Mode on page 40).

```
java -jar Notung-2.9.jar -g "C:\Users\Notung\Desktop\Trees\GENETREE.nwk" -s
 "C:\Users\Notung\Desktop\Trees\SPECIESTREE.nwk" --rearrange
--speciestag postfix  --threshold 85% --edgeweights length
--treeoutput nhx --log
```

This command rearranges the gene tree in `GENETREE.nwk`, where the species are in postfix format, with the DL event model. The `--edgeweights length` flag indicates that edge weights in the gene tree are stored in the branch length location in the Newick file (see Section A.6 - Location of Edge Weight Values on page 131). The user specified rearrangement threshold is 85% of the maximum edgeweight in the tree. The output will be saved to `GENETREE.nwk.rearranged.0.nhx`. Notung diagnostics will be saved to the file `GENETREE.nwk.rearranged.ntglog.txt` in nhx format (see ch. 7).

```
java -jar Notung-2.9.jar -g GENETREE.nhx -s SPECIESTREE.nhx --root
--speciestag nhx --infertransfers true --treeoutput newick
```

This command performs DTL rerooting on `GENETREE.nwk` informed by `SPECIESTREE.nhx`. Species names in the gene tree are stored in the NHX species tag field in the gene tree file. The top two best-scoring rooted trees will be saved in Newick format in the files `GENETREE.nhx.rooted.0.nwk` and `GENETREE.nhx.rooted.1.nwk` (See Chapter 6 - Rooting Mode on page 68).

```
java -jar Notung-2.9.jar -b "C:\Users\Notung\Documents\batch\batch.run"
--speciestag prefix --rearrange --edgeweights name --threshold 65 --log
```

This command rearranges the gene tree(s) specified in batch.run, which all have the species names stored in prefix form and edgeweights in name form. The species tree is specified in the first line of `batch.run`. The user specified rearrangement threshold is 65. Diagnostic messages will be saved in the file `batch.run.rearranged.ntglog` (see Section 12.3 - Running Notung from a Batch File on page 105).

> *NOTE*: By default, file paths in the batch file will be interpreted as relative to the location of the batch file; with the `--absfilenames` flag, file paths are interpreted as absolute file names.

Appendix A

File Formats

Notung can save trees in three different file formats: **Newick file format, NHX file format**, and **Notung file format**.

Newick file format specifies tree topology and node labels, but cannot be used to save reconciliation information or information about the species tree with which the gene tree was reconciled.

NHX and Notung file formats use the Newick comment field to store additional information not captured in the standard Newick specification. A reconciliation involves a gene tree, a species tree, the mapping from gene tree to species tree, and the inferred events. NHX format can store a gene tree, with additional information to indicate which nodes are duplications. Notung file format can store a gene tree, the species tree with which it was reconciled, duplication and loss nodes, and transfer edges with donor and recipient species identified. If you save a reconciled tree in Notung format, it will display all information stored with the tree when it is next opened in Notung.

The Notung file format holds more information, but may not be compatible with other software packages that use Newick format. The formal specification of Newick file format allows bracket-delimited comments. Programs that follow the formal specification and ignore information stored in comments will be able to read NHX or Notung format trees. However, not all programs allow comments. If you plan to use a program that does not allow Newick comments to further analyze trees saved by Notung, save your trees in standard Newick format.

A.1 Newick File Format

Newick is widely used by phylogeny programs. PHYLIP [11], PAUP* [26], and many other programs will output trees in Newick.

The general Newick syntax looks like this:

```
treefile → subtree;
subtree → descendant_list [internal_node_label] [:branch_length]
descendant_list → (subtree, subtree [, subtree]) | leaf_node_name
```

where `descendant_list` is a string that specifies the organization of the subtree and `internal_node_label` is the label of the root of a subtree. The optional `branch_length` field refers to the length of the edge from the root of the subtree to its parent. The `internal_node_label` and `branch_length` fields are optional. Some programs use these fields to store other information. For example, Notung allows the user to use either of these fields to store edge weight values.

Comments in Newick format are enclosed in square brackets and may appear anywhere newlines are permitted. Some programs use the comment field to store additional information that is not included in the Newick specification. By convention, this information is formatted as follows:

```
[&&ApplicationID:Application_specific_comments]
```

where `ApplicationID` indicates a specific program or format.

For more information about Newick file format, go to:

http://evolution.genetics.washington.edu/phylip/newicktree.html.

or

http://geta.life.uiuc.edu/~gary/Newicks_845_Tree_Std.html.

A.2 NHX File Format - New Hampshire eXtended

NHX File Format is based on the Newick file format, but embeds additional information about each node in the tree in the comment fields, as follows:

```
[&&NHX:TagID1=value1:TagID2=value2]
```

where `TagID1` and `TagID2` can specify bootstrap values, species labels, or event information. This example has two tags, but NHX comments can have one or more tags. Trees saved in NHX file format include information produced by a reconciliation, including duplications, co-divergences, transfer, and species labels, but do not record any visual annotations made in Notung. Nor do they record the species tree with which the gene tree was reconciled.

NOTE: The NHX format is case-sensitive.

More information about NHX format, including a complete list of tags used in comment fields, can be obtained at:

http://www.phylosoft.org/NHX/.

A.3 Notung File Format

Notung File Format further extends the NHX format. Notung file format can record duplication marks, transfer edges, edge weights, and color annotations. A reconciled gene tree file saved in Notung format will also have the associated species tree embedded in it. When the reconciled gene tree is reopened in Notung, the species tree can be extracted and used in the same way as any other species tree. A reconciled gene tree saved in Notung file format also stores additional information on parameter values, including edge weight threshold, loss cost, duplication cost, co-divergence cost, and transfer cost. In addition, a *non-binary gene tree reconciled with a binary species tree* with more than one optimal history stores information regarding which history was displayed when saved. When the gene tree is reopened in Notung, the tree for that optimal history will be displayed.

To open an embedded species tree in a Notung format gene tree file:

1. Open the Notung format gene tree file.

2. Click the Reconciliation tab to enter reconciliation mode.

3. Click the "**Show Pruned Species Tree**" button.

NOTE: None of the three file formats used in Notung embed alternate histories for gene trees. Notung saves only the history that currently appears in the tree panel. To access the other alternate histories when opening such a file, the tree must be analyzed again in Notung.

A.4 Specifying the Species Associated with Each Gene

In order to perform reconciliation, Notung must determine the species from which each leaf taxon in the gene tree was derived. This is achieved by embedding the species name in the gene leaf label or by using information embedded in the NHX comment field.

Notung offers three different conventions for specifying the gene to species mapping, described below. Notung will attempt to guess the naming convention used; you can also specify this in the reconciliation dialog (see Chapter 5 - Reconciliation Mode on page 40).

- **Prefix**: When using this format, Notung checks each gene name to see if it begins with a label in the designated species tree. For example, the gene names `HuGST002`, `Mm_GST_5`, and `Cow_gene1` could correspond to species labels Hu, Mm, and Cow, respectively.

  > *NOTE*: When using this format, no species label should be a prefix of another species label, such as with `carp` and `carpinusBetulus`. In this situation, Notung may incorrectly identify the gene `carpinusBetulus_gene1` as a carp gene, rather than a hornbeam gene.

- **Postfix**: When using this format, Notung looks for underscores (_) in gene names, and takes the species name as everything to the right of the last underscore. For example, `gene1_HUMAN` and `gene3_COW` would be labeled as Human and Cow genes respectively. Notung can distinguish between species names such as `carp` and `carpinusBetulus` in this format.

  > *NOTE*: Postfix mode *cannot* be used if species names include underscores (_); for example, `Carpinus_betulus` cannot be used in Postfix mode.

- **NHX**: When using this format, Notung finds the species label in an NHX tag after the gene name. For example, the gene `gene2[&&NHX:S=human]` would be labeled as a human gene. In this case, the species name is not necessarily visible in the gene tree leaf label, but can be viewed by selecting the "**Display Leaf Node Species Name**"

option in the Display Options menu. Gene trees output by Notung in NHX or Notung file format are saved using the NHX species tag, even if the original tree used Prefix or Postfix modes. Notung can distinguish between species names such as `carp` and `carpinusBetulus` in this format.

A.5 Punctuation in Species Names

In previous versions of Notung, punctuation (`-`, `/`, `_`, `.`, `\`) in species names was used to indicate that Notung should look for a shorter species tag in gene names, rather than looking for the entire species name. For example, given the species name `Hu.Homo_Sapiens`, Notung would look for the species label "`Hu`" in gene names.

Because many users found this confusing, this functionality has been removed since Notung 2.6. Notung now looks for entire species names during reconciliation, which also allows users to use species names like `Pan_troglodytes` and `Pan_paniscus` in the same tree without creating a conflict. Unfortunately, this means that some trees that were used in previous versions of Notung will not work in the current version. This section explains how to change these trees so that they can be used with Notung 2.7.

How do I tell if I need to convert my trees?

Any species tree with punctuation in the species names, where the *full species names* are not present in either the gene tree names or in NHX style species tags, will need to be converted. If your species names contain punctuation and you used them with older versions of Notung, then your trees probably fit this description. If Notung 2.7 is used to open an older Notung format tree that needs to be converted, a warning dialog will be shown.

Converting the trees

There are three ways to convert trees with punctuation in species names. The correct method to use depends on your desired outcome.

Shorten species names This method requires changing only the species tree - gene trees should not need to be modified. Remove any part of the species name after the first punctuation, including the first punctuation. For example, if the leaf labels in the gene tree are of the form "`Hu-gene01`", change "`Hu.Homo_sapiens`" to "`Hu`" in the species

tree. These shorter species names should now match the species labels in the gene names.

Lengthen gene names This method requires changing the gene tree(s). Replace short species labels in gene names with full species names. For example, change "`Hu-gene01`" to "`Hu.Homo_sapiens-gene01`" in the gene tree. This solution will not work in Postfix mode if your species names contain underscores (_).

Add NHX style species tags This method requires changing the gene trees, but does not change gene names. One benefit of this method is that switching from a very short species label to a long species label will not affect the length of gene names.

If the gene tree is already in NHX or Notung format, modify the NHX comment after each gene name. To modify an existing NHX comment, find the species tag and replace the shorter species label with the full species name. For example, "`[&&NHX:S=Hu]`" becomes "`[&&NHX:S=Hu.Homo_sapiens]`".

If there are no comments in the file (i.e., the tree is in Newick format), add the following after each gene name: "`[&&NHX:S=<speciesname>]`", where `<speciesname>` is the corresponding full species name from the species tree. For example, the gene tree:

```
(gene1_Hu,
 (gene2_Hu, gene2_Mu));
```

would become:

```
(gene1_Hu[&&NHX:S=Hu.Homo_sapiens],
 (gene2_Hu[&&NHX:S=Hu.Homo_sapiens], gene2_Mu[&&NHX:S=Mu.Mus_musculus]));
```

A.6 Location of Edge Weight Values

Notung uses edge weights to determine which edges are weakly supported and may be rearranged. These edge weights may correspond to bootstrap values, probabilities, branch lengths, or any other numerical indication of support.

Edge weight values can be located in one of three places in a tree file, depending on how the file was created. In Newick format, either the branch length field or the internal node name may be used to specify edge weights. Many programs store bootstrap values in the Newick node name field. In an NHX or Notung format file, edge weights can also be specified using the NHX bootstrap tag in the comment field.

The example below shows a tree with a single edge weight in each of the three tree formats:

- Branch length field in Newick format:

```
(cow_gene1, (mouse_gene2, cow_gene2):100)
```

- Internal node name field in Newick format:

```
(cow_gene1, (mouse_gene2, cow_gene2)100)
```

- NHX bootstrap tag in the comment field:

```
(cow_gene1, (mouse_gene2, cow_gene2)[&&NHX:B=100])
```

Confusion can arise if an input tree has edge weights in more than one type of field. This could occur, for example, in a tree that has both branch lengths and bootstrap values. Notung tries to guess the type of edge weight specification in the file, but it is not always possible for Notung to determine this unequivocally. You can specify the location explicitly using command line options (see Chapter 12 - Command Line Options and Batch Processing on page 98) or using the "Select Location of Edge Weights" dialog in the Display Options menu (see Figure A.1).

To set the location of edge weights in Notung:

1. Click "**Display Options → Select Location of Edge Weights**." A dialog box appears.
2. Select one of the radio buttons (see Figure A.1).

   The gene tree will immediately reflect the change, so you can check the tree panel to verify that the choice you selected gives the desired values.
3. Click "**Apply**."

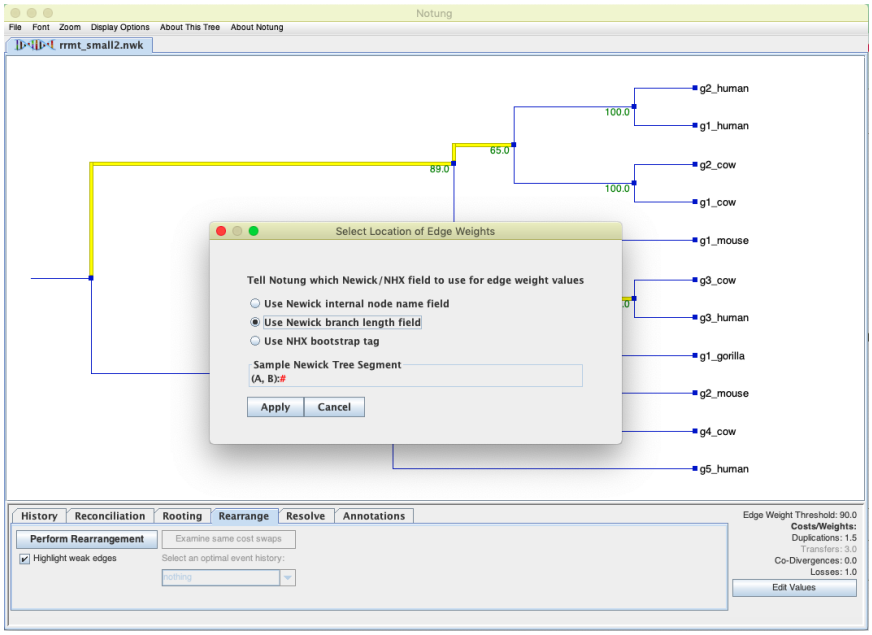


Figure A.1: The "Select Location of Edge Weights" dialog box.

Appendix B

Building a Species Tree

Most functions in Notung require a species tree. If you are familiar with the species in your data set, you may already have an appropriate species tree. If you do not have one, you can construct one using resources available on the web.

One such resource is the NCBI Taxonomy Browser, available at the NCBI website:

http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi

The Taxonomy Browser contains a database of all organisms represented in the NCBI sequence database, and can automatically build a species tree using species selected by the user. To create a tree in a format Notung can understand, add the species to be included in the tree, and then use the Taxonomy Browser's "Save As" option to save the tree as a Phylip tree. The Phylip option causes the tree to be saved in a variant of Newick format. The resulting tree can then be loaded into Notung as a species tree.

NOTE: The Taxonomy Browser does not recognize all common species names. Formal names for species can be found at:

http://www.expasy.org/cgi-bin/speclist

To build a species tree using the NCBI Taxonomy Browser:

1. Go to: http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi.
2. In the text field labeled "Enter name or id," enter the Latin name or common name of the species to add to the tree.

3. Click "**Add**." If the taxonomy browser did not recognize the species name, a pink error bar which reads "Organism name '*name*' not found" will appear.
4. When you have finished adding species, find the pull-down menu that says "text tree." Drag down and select "**phylip tree**."
5. Click "**Save As**," and save the species tree.

Additional resources provide access to existing species trees built by other researchers. TreeBASE (http://www.treebase.org/treebase/search.html) allows users to search for species trees from a large database of published papers. The Angiosperm Phylogeny Website and the Phylomatic Project provide species trees for plant species.

http://www.mobot.org/MOBOT/research/APweb/welcome.html

http://www.phylodiversity.net/phylomatic/phylomatic.html

Other tree-building tools are listed on Felsenstein's Phylogeny Programs website:

http://evolution.genetics.washington.edu/phylip/software.html.

NOTE:

- Species trees obtained from the Taxonomy Browser and other resources will not necessarily be bifurcating. To correct this, you can either edit the tree file in a text editor or tree editor or use the non-binary species tree with a binary gene tree. See Chapter 4.2 - Non-Binary Trees on page 29 for more information.
- The Phylip tree files generated by the Taxonomy Browser contain non-ASCII characters and have branch lengths of four within the tree, but these characters have no effect on the tree when it is opened in Notung.

Appendix C

Glossary

Binary Tree: A tree in which every internal node has degree three. If the tree is rooted, every internal node has one parent and two children. Also known as a **Bifurcating Tree**.

Bipartition: In the phylogenetic context, the separation of the leaf taxa into two sets. Each edge in the tree specifies the taxon bipartition that would result if the edge were removed. Also called a **Split**.

Combined Loss: The interpretation that the absence of a gene in two or more sibling species is evidence of a single loss in their common ancestor.

Co-Divergence: In a gene tree reconciled with a non-binary species tree, a node whose incongruence with the species tree could be due to either deep coalescence or duplication.

Co-Divergence Cost: The weight c_C of each co-divergence in the DTL Score. By default, $c_C = 0$.

Duplication/Transfer/Loss Score: The weighted sum, $c_L L + c_D D + c_C C + c_T T$, of losses (L), duplications (D) co-divergences (C), and transfers (T) in a reconciled gene tree. Also known as the DTL Score or the **D/T/L cost**. If transfers are not inferred, it is the DuplicationLoss or DL Score.

Deep Coalescence: The divergence of genes when the time of separation of a gene lineage predates the time of speciation.

Duplication Cost: The weight c_D of each duplication in the DTL Score. By default, $c_D = 1.5$

Edge Weight: A numerical value representing a quantitative assessment of the support in the underlying data for the associated bipartition. Typically bootstrap values, branch

lengths, or likelihood scores are also used as edge weights. If used, edge weights are specified in the input gene tree file.

Edge Weight Threshold: Numerical value used to define strong edges. Edges with weights below the edge weight threshold are considered unreliable or weak.

Event History: A set of event-edge pairs, where each event is a duplication, co-divergence, transfer or loss and each edge is an edge in the species tree that specifies when the event occurred. An event history specifies a set of trees in which any tree in the set can be obtained from any other tree in the set by a series of **Same Cost Swaps**.

Hard Polytomy: A polytomy that represents the simultaneous divergence of the three or more lineages. Hard polytomies are only found in species trees. See also **Soft Polytomy**.

Height: The maximum path length between any leaf and the root of a tree.

Incomplete Lineage Sorting: Incongruence between a gene and species tree that occurs when the lineages of a gene tree sort independently from the lineages of the associated species tree.

Loss Cost: The weight c_L of each loss in the DTL Score. By default, $c_L = 1.0$

Non-binary Tree: A tree in which at least one node has degree greater than three. In a rooted non-binary tree, at least one node has more than two children. Also known as a **Multifurcating Tree**.

Orthologs: Homologous genes whose LCA is found in the cenancestor of the species from which the genes were sequenced.

Paralogs: Homologous genes whose LCA is associated with a duplication event.

Paraxenologs: Homologous genes that are both paralogs and xenologs.

Polytomy: A node with degree greater than three. In a rooted tree, a node with more than two children.

Polytomy Size: The number of children of a polytomy in a rooted tree.

Pruned Species Tree: A species tree containing only the species that appear in the gene tree with which it was reconciled.

Rearranged Tree: A reconciled, binary gene tree with minimal D/T/L Score that agrees with the original tree at all strongly supported edges.

Reconciled Tree: A gene tree that has been fit to a species tree, resulting in a mapping between each node in the gene tree and a node in the species tree. From this mapping, gene losses, duplications, co-divergences, and transfers are inferred.

Resolved Tree: A binary tree derived from a non-binary gene tree, in which each polytomy has been removed and replaced with a set of binary divergences.

Same Cost Swap: An interchange of two nodes in the gene tree that does not change the DTL Score of the tree, and does not break any **Strong Edges**.

Soft Polytomy: A polytomy that represents uncertainty in the true, binary branching pattern of its descendant lineages. Soft polytomies can be found in either gene or species trees.

Strong Edge: An edge with weight greater than or equal to the **Edge Weight Threshold**. Any edge without a specified weight is assumed to be weak.

Transfer Cost: The weight c_T of each transfer in the DTL score. By default, $c_T = 3.0$.

Trifurcation: In a rooted tree, a node with exactly three children.

Weak Edge: An edge with weight lower than the **Edge Weight Threshold**. Any edge without a specified weight is assumed to be weak.

Xenologs: Homologous genes who have undergone a horizontal gene transfer event since their LCA.

Appendix D

Keystroke Shortcuts

Key Combination	**Action**
Ctrl + O	Open a gene tree
Ctrl + Shift + O	Open a species tree
Ctrl + S	Save the tree
Ctrl + P	Print the current view
Ctrl + Shift + R	Reload tree from file
Ctrl + W	Close tree
Ctrl + =	Increase font size (for all labels in the tree)
Ctrl + -	Decrease font size (for all labels in the tree)
Ctrl + *click on tree*	Zoom in on tree
Shift + *click on tree*	Zoom out of tree
Ctrl +]	Zoom in on tree on the X-axis
Ctrl + [	Zoom out of tree on the X-axis
Ctrl + Shift +]	Zoom in on tree on the Y-axis
Ctrl + Shift + [	Zoom out of tree on the Y-axis
Ctrl + T	Show whole tree
Ctrl + .	Go to next tree
Ctrl + ,	Go to previous tree
Ctrl + Q	Exit (end Notung)

NOTE: Ctrl indicates use of the control key. Ctrl + *click on tree* means that the user needs to click on the tree while pressing the appropriate key. Mac users may have to use the *command*, or *open apple* key to zoom in on the tree (*i.e.*, command + *click on tree*), but should use the control key for all other operations.

Appendix E

Worked Examples

The following exercises will help familiarize you with the basic tasks Notung can perform on a gene tree. The tree files used in these exercises are included in the Notung distribution, in the *sampleTrees* folder. If the program window becomes too cluttered, you may close trees that are no longer being used by selecting the tree and clicking on "**File** → **Close**."

E.1 Exercise 1 - Reconciling a gene tree with a species tree

In this exercise, you will reconcile the gene tree `genetree_NOTCH` with the species tree `speciestree_mega`. You will also generate a pruned species tree, and use Notung to determine the upper and lower bounds on the time when a duplication occurred.

Open the tree files

1. Click "**File** → **Open Gene Tree**" and open `genetree_NOTCH`.

   The gene tree is located in the *sampleTrees* folder, which is included in the downloaded zip file. Once loaded, the gene tree is displayed in the tree panel.

2. Click "**File** → **Open Species Tree**" and open `speciestree_mega`.

   The species tree is located in the *sampleTrees* folder. Once loaded, the species tree appears in the tree panel. Because it is the most recent tree opened, it is now selected.

Note that the options that Notung offers differ depending on whether a species tree or a gene

tree is selected. For example, because `speciestree_mega` is now selected, the box showing parameter values in the lower right corner has disappeared, and the task panel includes only two task modes, **History** and **Annotations**.

Reconcile the gene tree with the species tree

1. Click on the `genetree_NOTCH` tab to select the gene tree.

2. Click the "**Reconciliation**" tab.

   The **Reconciliation** task panel opens below. From here you can reconcile a gene tree with a species tree, display a pruned species tree, show duplication bounds, and hide duplication marks and loss nodes.

3. Check the "Prune" box. The leaves of `speciestree_mega` include more species than are relevant to `genetree_NOTCH`. These species are removed from the tree when this checkbox is selected.

4. Click "**Reconcile/Rereconcile**."

   The **Reconciliation** dialog appears. In this dialog box, Notung asks you to specify which species tree to use for the reconciliation and what naming convention is used in the gene tree to specify the species associated with each gene.

5. Select `speciestree_mega` in the drop-down menu labeled "Please select a species tree to reconcile with."

   Currently, the only selection available is `speciestree_mega`. However, if you have more than one species tree open in Notung, you must specify here which species tree to use.

6. Under the section labeled "Specify Species Label" select "**Prefix of the gene label**."

   This section in the dialog box asks you to specify the naming convention used in the gene tree to indicate from which species the genes originated. Notung tries to guess the naming convention, but it does not always guess correctly. Notung should have guessed correctly in this case. In general, remember to check the leaf node names in your gene tree during this step to make sure that they agree with the naming convention you choose.

   For more details about the species label naming conventions, see Appendix A.4 - Specifying the Species Associated with Each Gene on page 129.

7. In the dialog box, click "**Reconcile**."

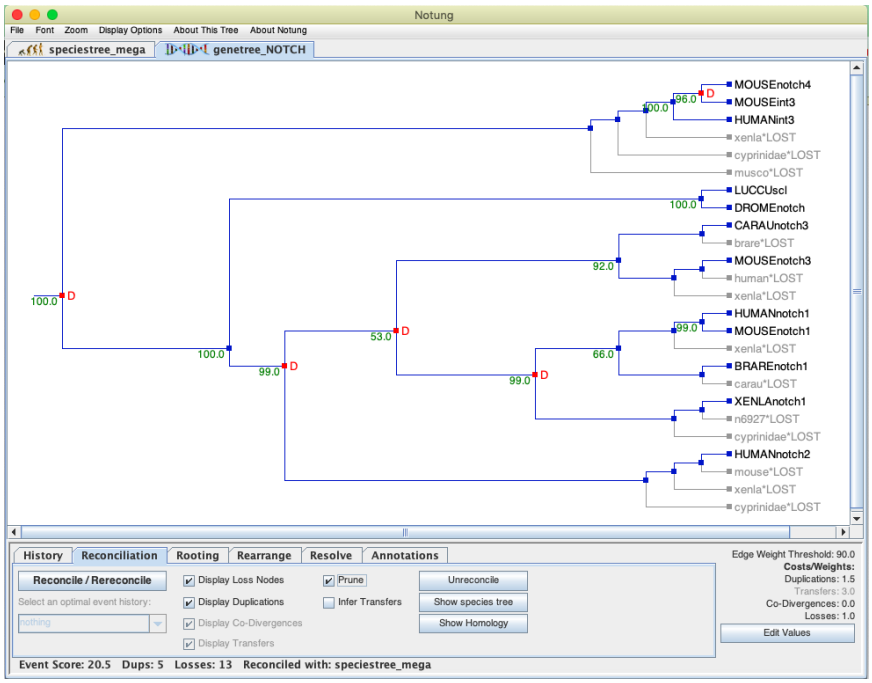


Figure E.1: The gene tree should now look like this.

The reconciled gene tree now appears in the tree panel. The DL Score of the reconciled tree, displayed in the bottom-left corner of the program window, is 20.5 - five duplications and thirteen losses. Five red D's in the tree indicate the inferred duplications. At the right end of the tree (at the leaves), thirteen loss nodes appear in light gray type.

Display the pruned species tree

After reconciliation, you can view the associated species tree. If the Prune checkbox is checked, the species tree will be pruned of all species that are not represented by genes in the gene tree.

1. Click the "**Show Pruned Species Tree**" button.

   A dialog box appears asking you to give a title for the pruned species tree. The default title is "Pruned Species Tree."

2. In the dialog box text field, enter "Mega_Pruned" (or any other name you like), then click "**OK**."

   The pruned species tree appears in the tree panel. It contains only seven leaf nodes, all of which are species represented in the reconciled gene tree. The pruned species tree has a tab above the tree panel, labeled "Mega_Pruned." You can now select and use this tree as you would any other species tree.

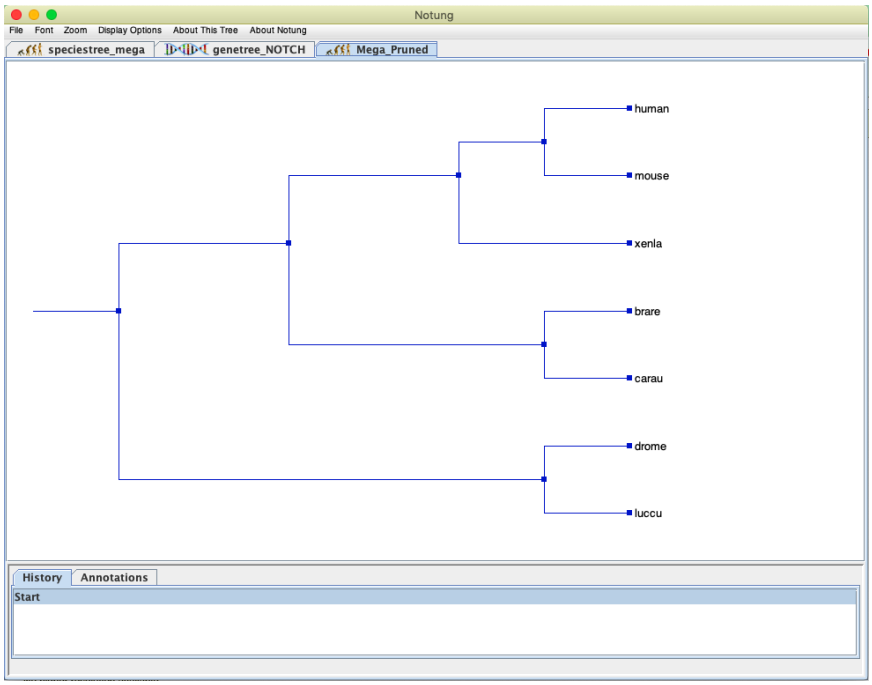


Figure E.2: This is what you should see after reconciling and then clicking "show pruned species tree."

Check the event details

The event summary provides information regarding when events occurred in the course of species evolution.

1. Select `genetree_NOTCH`.
2. Click "**Display Options → Display Internal Node Names**."

   Node name labels appear in red type next to each internal node. You can now identify each duplication by name. If internal node names are not provided in the gene tree file, Notung will assign the node an alphanumeric name (*e.g.* `n132`).
3. Select the `Mega_Pruned` species tree.
4. Click "**Display Options → Display Internal Node Names**."

   Node name labels appear in red type next to each internal node.
5. Select `genetree_NOTCH`.
6. Click the "**Reconciliation**" tab.
7. Click on "**About This Tree → Event Summary**" menu item.

A new window appears. Inferred duplications are listed in the left column, expressed as node names in the gene tree. The lower and upper bounds are listed in the middle and right columns, respectively, and are expressed as internal node names in the species tree. Information on losses is displayed below duplication bounds. The left column lists the species nodes in the species tree. The right column provides the number of losses that occurred in each species.

8. Find the duplication node in the bottom-right area of the tree, from which the `XENLAnotch1` gene extends. Find its node name and duplication bounds in the "Duplications and Losses" window.

   The node name may vary, depending on how many internal nodes Notung has counted in your current session.

9. Close the window and select the pruned species tree `Mega_Pruned`.

   With `Mega_Pruned` selected, you can see internal nodes representing *euteleostomi* and *coelom*. The duplication occurred somewhere on the edge between those nodes.

E.2 Exercise 2 - Rooting an unrooted tree

The gene tree `genetree_ANK` is unrooted. In this exercise, you will select a root based on duplication loss parsimony.

Open the tree files

1. Click "**File → Open Gene Tree**" and open `genetree_ANK`.

   The gene tree is located in the *sampleTrees* folder.

   Since this tree is unrooted, it has a trifurcation (a node with 3 children) at the top of the tree, but is otherwise binary.

2. Click "**File → Open Species Tree**" and open `speciestree_mega`. If `speciestree_mega` is already opened, you may skip this step.

3. Be sure that the `genetree_ANK` tab is selected before proceeding.

Run the Rooting Analysis

1. Click the "**Reconciliation**" tab and be sure the "Prune" option is checked. This will ensure that our species tree is pruned before it is reconciled with this gene tree.

2. Click the "**Rooting**" tab.

   The **Rooting** task panel is displayed. Notung is now in **Rooting** mode.

3. Click "**Run Rooting Analysis**."

   You will be asked to reconcile the tree. Select `speciestree_mega` and "Prefix," click "**Reconcile**". The edge at the top of the tree panel, leading to `caeel*unc-44`, is colored red. This means it has the minimum root score.

4. Optional: Click the "**Display root score**" checkbox.

   Each edge is labeled with its root score. Notice that the red edge leading to `caeel*unc-44` has a root score of 4.0. The next lowest score is 8.5.

Select a root

1. Click on the **red edge** in the tree panel.

   The tree is now rooted on the edge leading to the `caeel*unc-44` gene. The DL Score of the tree is now 4.0, with two duplications and one loss.

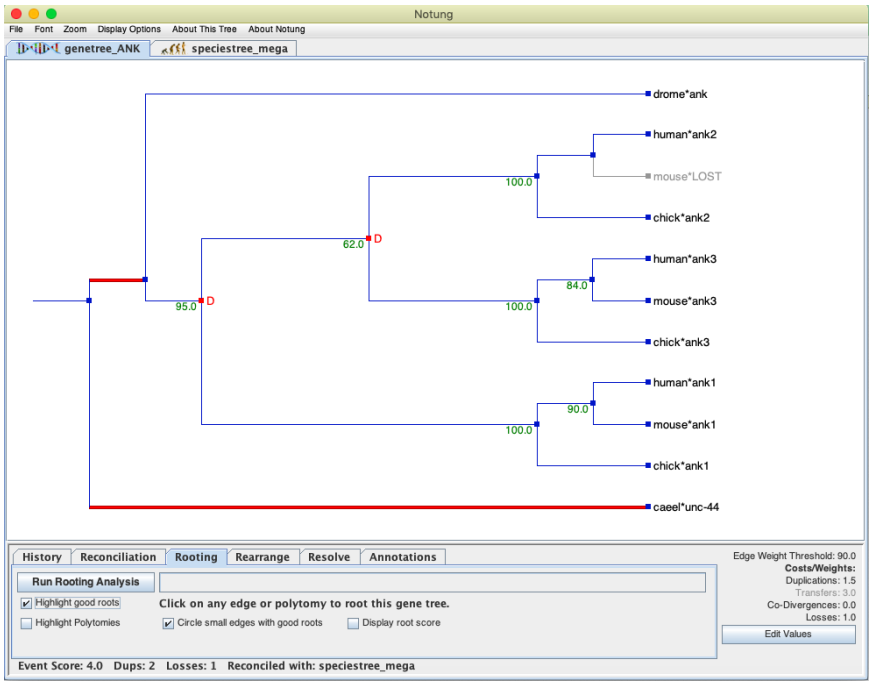


Figure E.3: The gene tree should now look like this.

E.3 Exercise 3 - Rearranging a gene tree

In this exercise, you will reconcile the gene tree `genetree_SMALL` with the species tree `speciestree_small` and use Notung's rearrangement tasks to investigate alternate gene trees with minimum DL Score. Both input trees are located in the *`sampleTrees`* folder.

Reconcile the gene tree with the species tree

1. Click "**File → Open Species Tree**" and open `speciestree_small`.
2. Click "**File → Open Gene Tree**" and open `genetree_SMALL`.

   This is an artificial tree made up for this exercise. The edge weights in this tree represent bootstrap values. Note that two internal edges have a bootstrap value of 100, one has a bootstrap value of 73, and several have not been assigned a weight. (Note that edges adjacent to leaves are usually not assigned bootstrap values since those edges are present in all trees.) Notung sets the default edge weight threshold to 90% of the maximum edge weight in the tree. Since the maximum edge weight in this tree is 100, the edge weight threshold is set to 90.0.
3. Click the "**Reconciliation**" tab.
4. Click "**Reconcile/Rereconcile**."
5. In the "Reconciliation Options" dialog box, select `speciestree_small` and "Postfix" and click "**Reconcile**."

   The reconciled tree appears in the tree panel. Note that it has a DL Score of 10.0, with four duplications and four losses.

Rearrange the reconciled tree

1. Click the "**Rearrange**" tab.

   The **Rearrange** task panel is now displayed.
2. Click the "**Highlight weak edges**" checkbox.

   Several edges in the reconciled tree are highlighted in yellow. These are edges with weights below the Edge Weight Threshold and are considered "weak." Weak edges may be rearranged to reduce the number of duplications and losses in the tree. Edges with weights above the threshold will not be rearranged.

   Note that in addition to the edge with weight 73.0, the internal edges with no edge weight are also highlighted in yellow. Notung assumes that any internal edge that is not explicitly assigned a weight is considered weak.

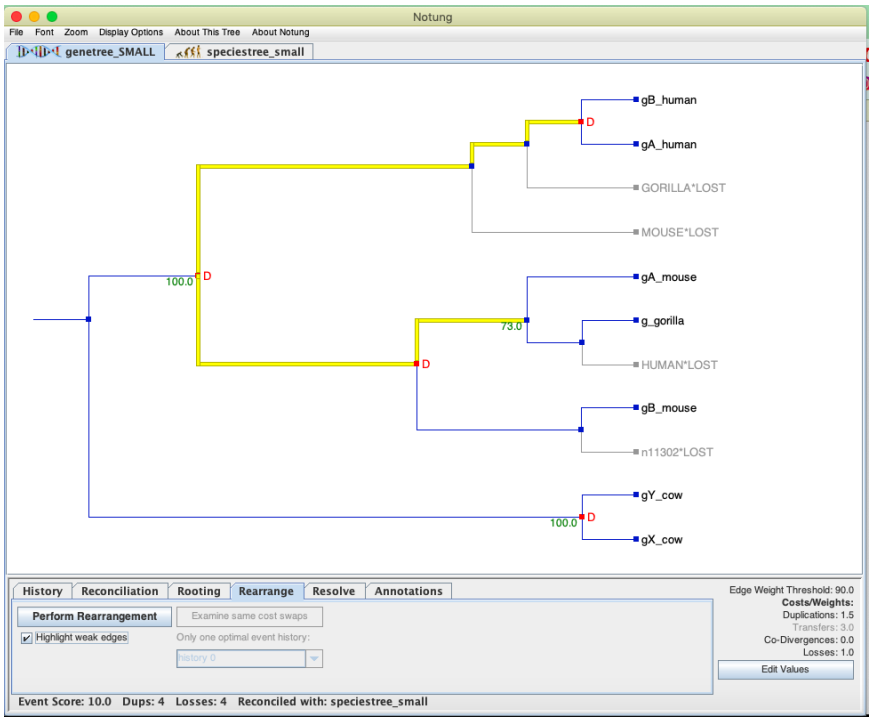


Figure E.4: The gene tree with weak edges highlighted.

3. Click "**Perform Rearrangement**."

The rearranged tree appears in the tree panel. It now has a D/L Score of 4.0, with two duplications and only one loss.

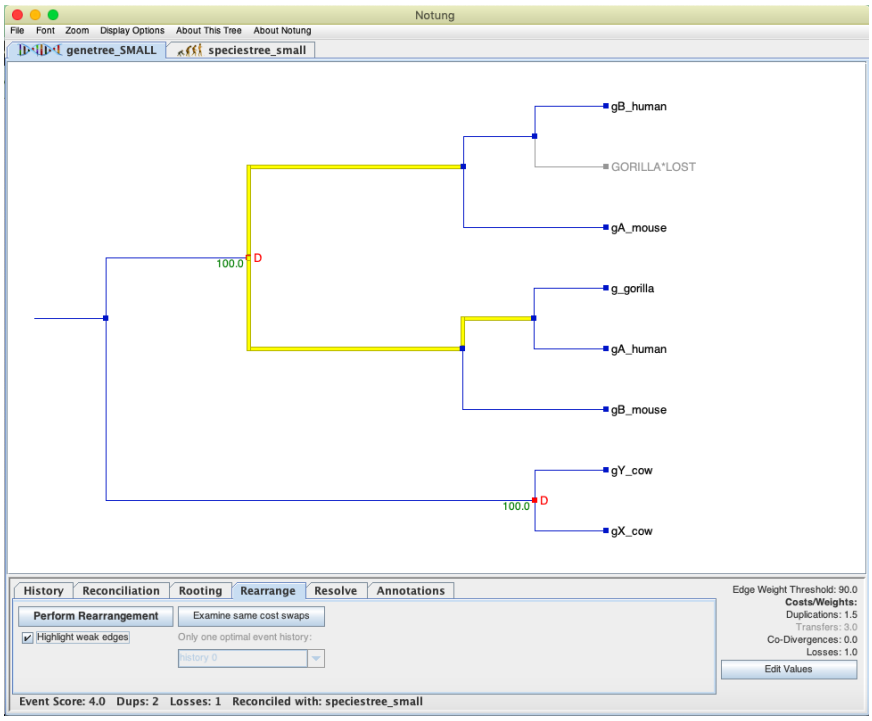


Figure E.5: The gene tree should now look like this.

Change the parameter values and rearrange again

In the previous steps, we rearranged the tree using the default parameter values ($c_D = 1.5$ and $c_L = 1.0$). For the default values, there is only one minimum cost tree. We now explore what happens when we rearrange the tree when duplications and losses are weighted equally.

1. Click the "**Edit Values**" button in the bottom-right corner of the program window.
2. In the dialog box, change the Duplication Cost to 1.0.
3. Click "**Apply Changes**."

   A message appears to warn us that although we have changed the parameter values, this has had no effect on the tree. We must rearrange the tree again to see the effect of rearrangement with this choice of parameter values. Click "**OK**."

   Duplications and losses are now weighted equally in Notung's reconciliation algorithm.
4. Click the "**Rearrange**" tab, if it is not already selected.
5. Click "**Perform Rearrangement**." The tree is rearranged with the new parameter values. The newly rearranged tree appears in the tree panel. The DL Score of this tree is 3.0, with three duplications and no losses.

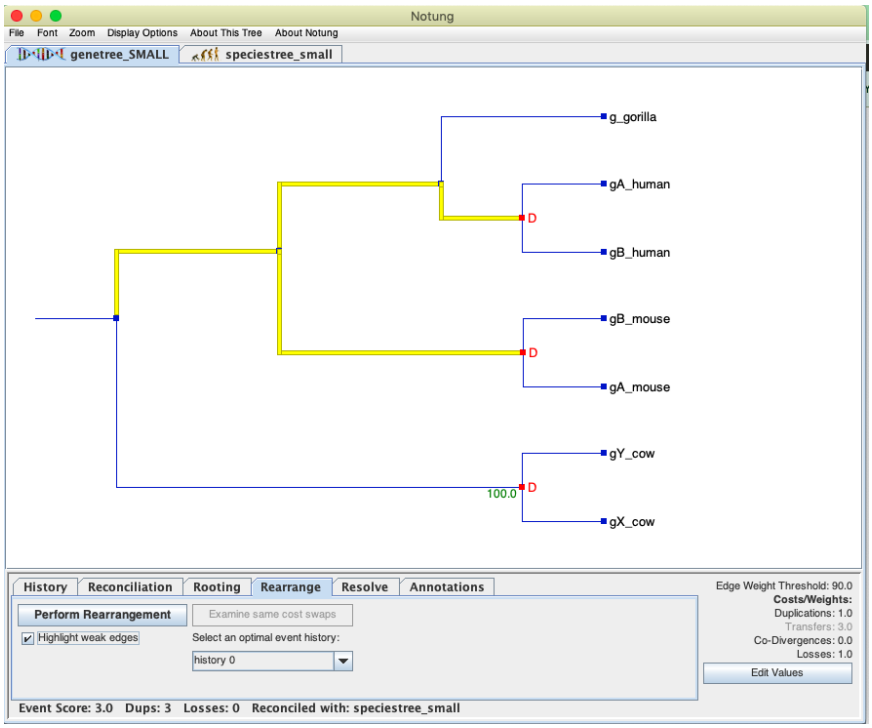


Figure E.6: The gene tree should now look like this.

View a different alternate event history

With the new parameter values, there is more than one alternate gene tree with minimal DL Score. You are currently viewing **history 0**.

1. In the **Rearrange** task panel, click on the drop-down menu labeled "Select an optimal event history."

   This opens a list of available alternate event histories. You should see `history 0` and `history 1`.

2. Select `history 1`.

   A different tree appears. This tree also has a D/L Score of 3.0, but has two duplications and one loss instead of three duplications and no losses.

Swap nodes in the rearranged tree

Note that this tree groups `gB_human` with `gA_mouse` and `gA_human` with `gB_mouse`. However, the tree that groups `gA_human` with `gA_mouse` and `gB_human` with `gB_mouse` has the same score.

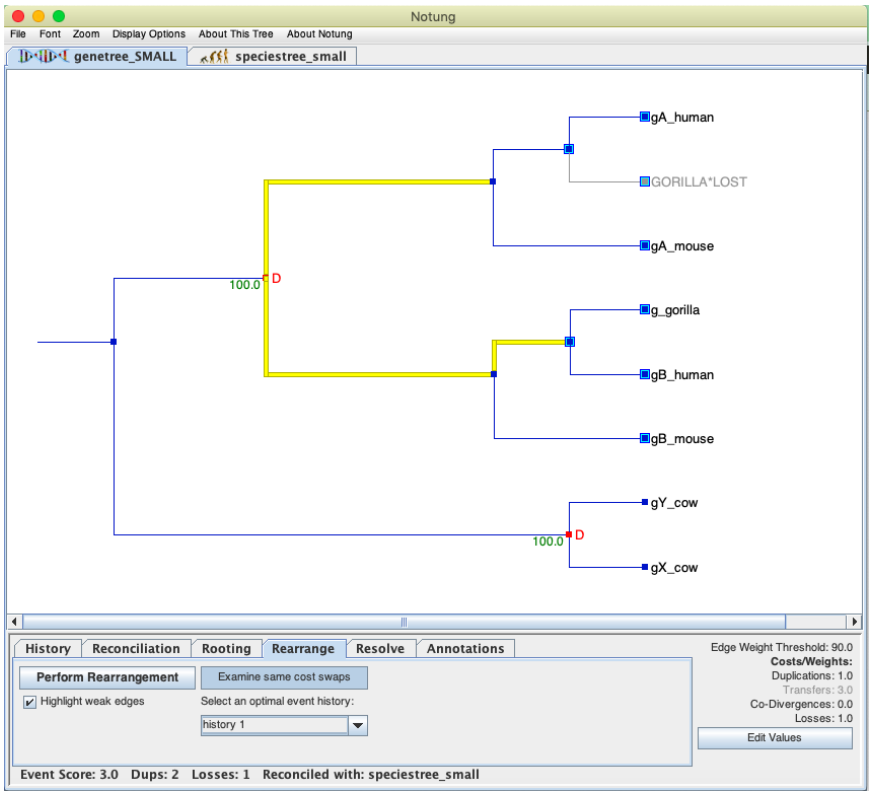


Figure E.7: The gene tree should now look like this.

1. Click the "**Examine same cost swaps**" button.

   Nodes that can be interchanged without changing the DL Score are marked with enlarged light blue boxes.

2. Click the `gB_human` node.

   To select the node, you must click on the enlarged blue box. When you click and select a node, a blue triangle will mark the node(s). Once selected, the node is marked with a light blue triangle. Each node it can be swapped with is marked with a pink triangle. In this case, there is just one: `gA_human`.

3. Click the `gA_human` node.

   The nodes `gB_human` and `gA_human` are swapped. Once they have been swapped, they are temporarily highlighted with yellow triangles, so that you can see the results of the most recent action. Note that the `gA` genes are now grouped together, and the `gB` genes are together in the same subtree, along with the `g_gorilla` gene.

   Try performing additional swaps to see how many alternate, minimum cost trees you can find.

E.4 Exercise 4 - Inferring duplications and losses in a binary gene tree with a non-binary species tree

In this exercise, you will perform Notung's main tasks to infer duplications and losses on the gene tree `exercise4_genetree` with the non-binary species tree `exercise4_speciestree`. You will reconcile and root the gene tree, and use Notung to determine the upper and lower bounds on the time when a duplication occurred.

Open the tree files

1. Click "**File** → **Open Gene Tree**" and open `exercise4_genetree`.

   This is an artificial tree made up for this exercise.

2. Click "**File** → **Open Species Tree**" and open `exercise4_speciestree`.

   As you will notice, this is a non-binary species tree with a polytomy representing the common ancestor of the marsupials.

Reconcile the gene tree with the species tree

1. Select `exercise4_genetree`.

2. Click the "**Reconciliation**" tab.

3. Click "**Reconcile/Rereconcile**."

4. In the "Reconciliation Options" dialog box, select `exercise4_speciestree` and "Prefix" and click "**Reconcile**."

   The reconciled tree appears in the tree panel. Note that it has a DL Score of 8.0, with two duplications, one co-divergence, and five losses. Two red D's in the tree mark the duplications, while the one pink cD marks the co-divergence. At the leaves of the tree, five loss nodes appear in light gray type.

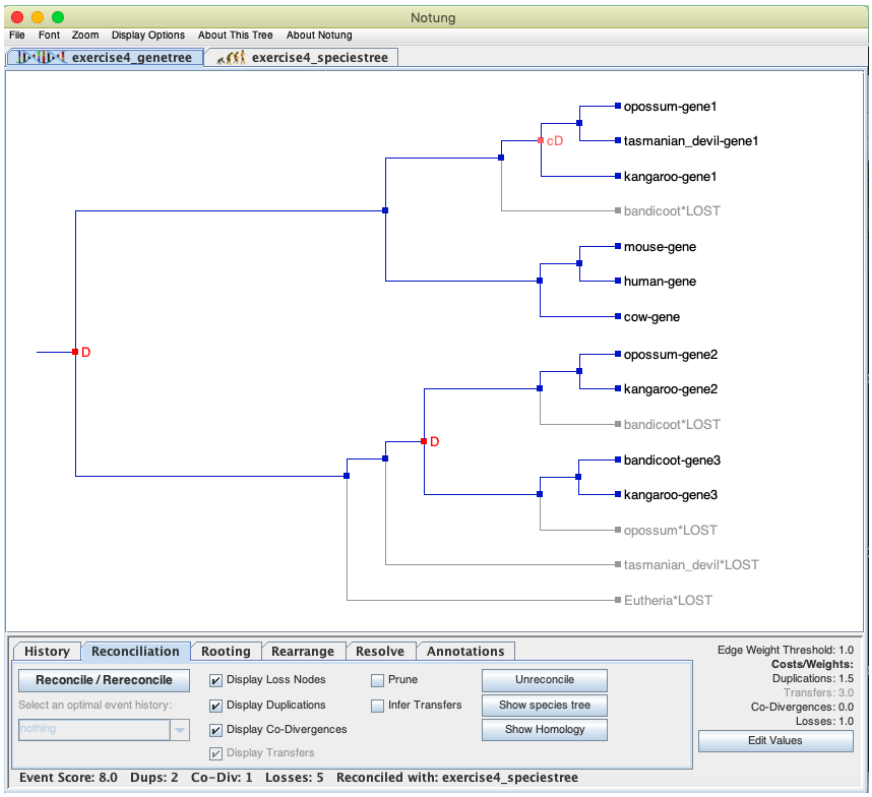


Figure E.8: The gene tree should now look like this.

Check the event summary

The event summary provides information regarding when events occurred in the course of species evolution.

1. Click "**Display Options → Display Internal Node Names**."
2. Select `exercise4_speciestree`.
3. Click "**Display Options → Display Internal Node Names**."
4. Select `exercise4_genetree`.
5. Click the "**About This Tree → Event Summary**" from the menu.

   In the new window, duplications are described first. Co-divergences are described below the duplications. For both types of events, the gene nodes are listed in the left column, expressed as node names in the gene tree. The lower and upper bounds are listed in the middle and right columns, respectively, and are expressed as internal node names in the species tree. Information on losses is provided below the co-divergence bounds.

6. Close the window by clicking the "**Close this window**" button.

Run the Rooting Analysis

1. Click the "**Rooting**" tab.

2. Click "**Run Rooting Analysis**."

   The edge leading to genes from placental mammals (cow, mouse, and human) is colored red. This means it has the lowest root score.

3. Under the section labeled "Specify Species Label" select "**Prefix of the gene label**."

4. In the dialog box, click "**Reconcile**."

5. Optional: Deselect "Display Internal Node Names" (under "Display Options") and click the "Display root score" checkbox.

   Notice that the red edge has a root score of 7.0. The next lowest root score is 8.0.

6. Click "**Display Options → Display Internal Node Species Names**."

   The name of the species to which the node is mapped appears in italics next to each internal node.

7. Select the optimal root by clicking on the **red edge** in the tree panel.

   The tree is rooted on the edge which splits the tree between placental mammals (*Eutheria*) and marsupials (*Metatheria*). The DL Score of the tree is now 7.0, with two duplications, one co-divergence, and four losses.

   Do not close these trees yet - they will be used in upcoming steps.

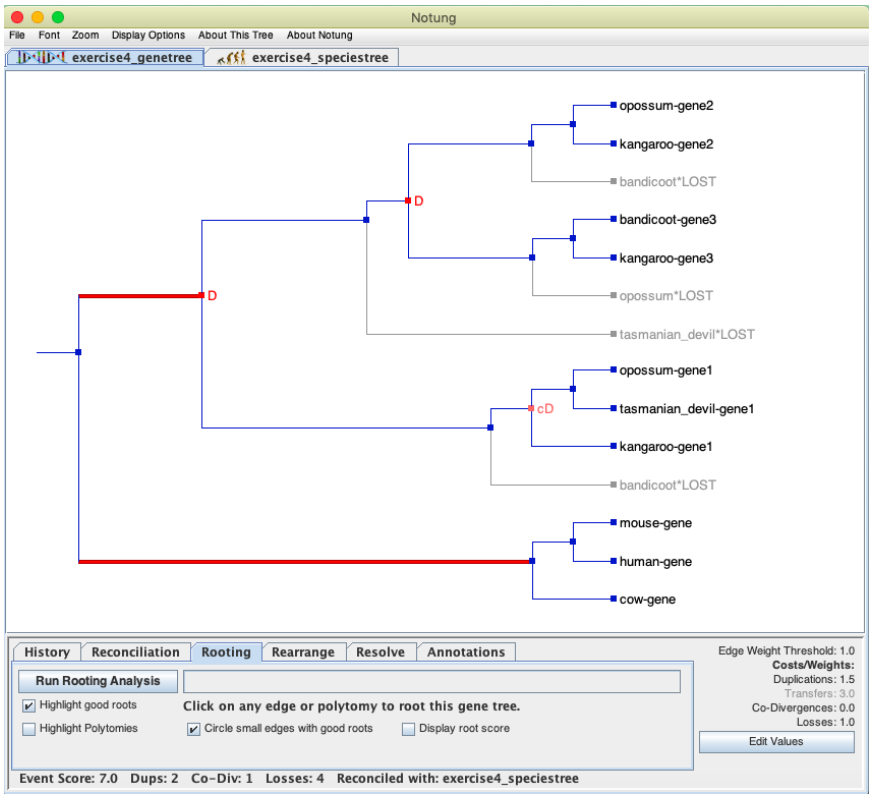


Figure E.9: The gene tree should now look like this.

Reconcile the Tree using the Combined Polytomy Losses algorithm

This step uses the command line interface and can be skipped, if desired. You will use the command line interface to reconcile the gene tree `exercise4_genetree` with the species tree `exercise4_speciestree` using the combined losses algorithm.

1. On the command line, navigate to the Notung directory.

   For instructions on using Notung from the command line, see Chapter 12.2 - Running Notung from the command line on page 102.

2. Type the following in the command window/terminal and hit enter:

   ```
   java -jar Notung-2.7.jar sampleTrees/exercise4_genetree -s
   sampleTrees/exercise4_speciestree --reconcile --exact-losses
   --outputdir sampleTrees --report-heuristic-losses
   ```

   Notung will print information to the screen as it reconciles the tree for both combined and explicit losses. Notice that the first unrooted gene tree has a DL Score of 8.0, with two duplications, one co-divergence and five heuristic losses as compared

to the second unrooted gene tree, which has a DL Score of 7.0, with two duplications, one co-divergence, and *four* exact losses. The tree, reconciled and with exact losses, will be saved to the *sampleTrees* folder (as specified by `--outputdir`) as `exercise4_genetree.reconciled`.

Root the tree reconciled with the Combined Polytomy Losses algorithm

In the previous step, you reconciled the gene tree while using the combined polytomy losses algorithm. In this step you are will find the optimal root for this gene tree. If you skipped the previous step, you will need to use the gene tree `exercise4_genetree-exactLosses.ntg` instead of `exercise4_genetree.reconciled`.

1. In Notung's graphical user interface, click "**File** → **Open Gene Tree**" and open `exercise4_genetree.reconciled`.

   If you skipped the last step, use `exercise4_genetree-exactLosses.ntg` instead.

   A warning will appear stating that the tree was reconciled using `--exact-losses`. Click the "**OK**" button.

2. Click the "**Reconciliation**" tab.

3. Click the "**Show pruned species tree**" button.

   A window will appear, and you can enter a title for the pruned species tree. Click the "**OK**" button.

   Switch back to the gene tree.

4. Click the "**Rooting**" tab.

5. Click "**Run Rooting Analysis**."

   A window will appear asking the position of species label. Choose the third option "**NHX species tag contains the species label**", and then click "**OK**"

6. Select the optimal root by clicking on the **red edge** in the tree panel.

   The tree is rooted on the edge leading to placental mammals. The DL Score of the tree is now 6.0, with two duplications, one co-divergence, and three losses.

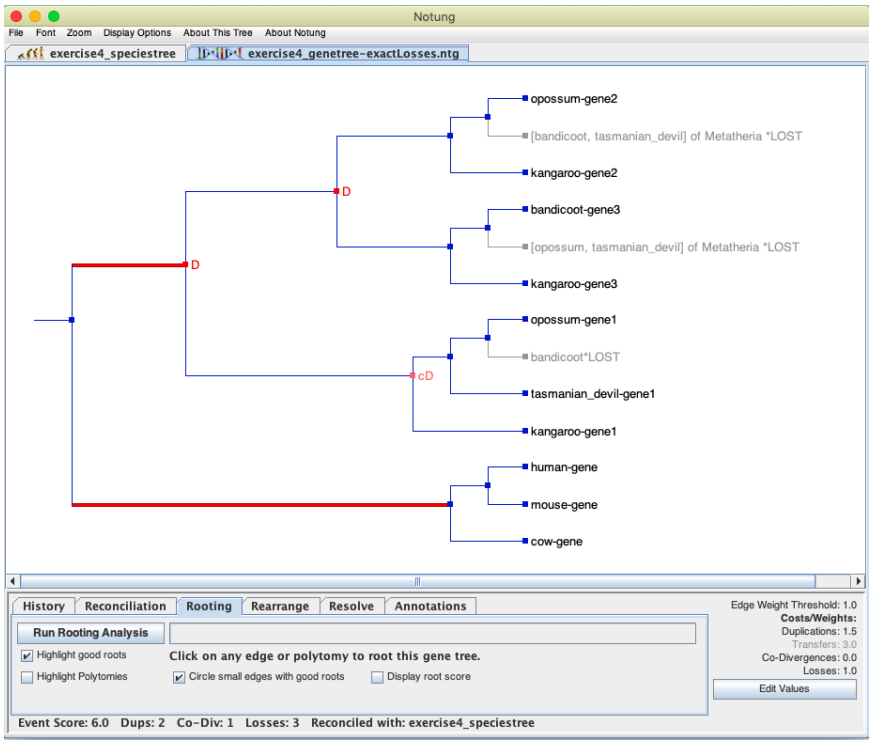


Figure E.10: The gene tree should now look like this.

Compare this tree with the previously rooted gene tree (`exercise4_genetree`). Can you find the difference between the trees? In `exercise4_genetree`, the loss node, `tasmanian_devil*LOST`, above the subtree containing genes `gene3` and `gene2`, has been moved below the duplication node and combined with `opossum*LOST` and `bandicoot*LOST` in the `gene3` and `gene2` subtrees, respectively, in `exercise4_genetree.reconciled`. This resulted in a reduction of the total number of losses.

View polytomy losses without species names included

There are two display options for polytomy losses. In this step, you will see the other way to display these losses.

1. Select the `exercise4_genetree.reconciled` gene tree. (Use `exercise4_genetree-exactLosses.ntg` if you skipped the step for reconciling the tree using the exact losses algorithm.)

2. Deselect the "**Display Options → Use Species Names in Polytomy Losses**" option. The gene tree no longer shows the species names for polytomy losses. For example the loss that was previously displayed as "`[tasmanian_devil, bandicoot] of Metatheria*LOST`" is now displayed as "`2/4 of Metatheria*LOST`."

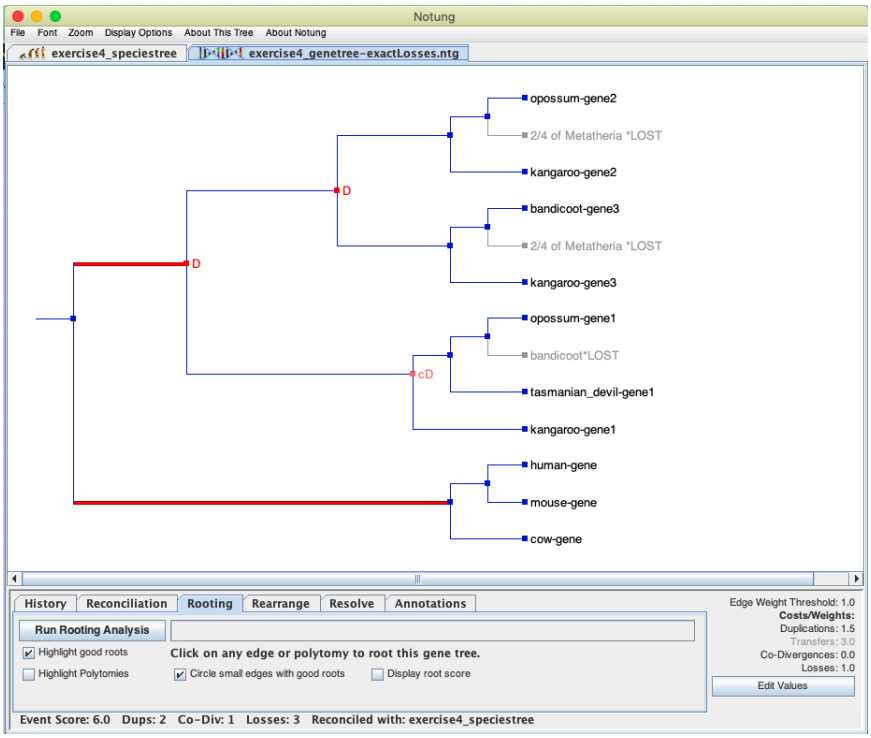


Figure E.11: The gene tree should now look like this.

E.5 Exercise 5 - Non-binary gene tree with a binary species tree

In this exercise, you will perform Notung's main tasks to infer duplications and losses on the non-binary gene tree `exercise5_genetree` with the species tree `exercise5_speciestree`. You will reconcile, root, resolve, and rearrange the gene tree, and use Notung to determine some general statistics about the trees.

Open the tree files

1. Click "**File** → **Open Gene Tree**" and open `exercise5_genetree`.

   This is an artificial tree made up for this exercise. Notice that this gene tree is non-binary and contains multiple polytomies.

2. Click "**File** → **Open Species Tree**" and open `exercise5_speciestree`.

3. Select `exercise5_genetree`.

4. Click "**Display Options → Highlight Polytomies**."

   The polytomies in the gene tree are circled and highlighted in cyan.

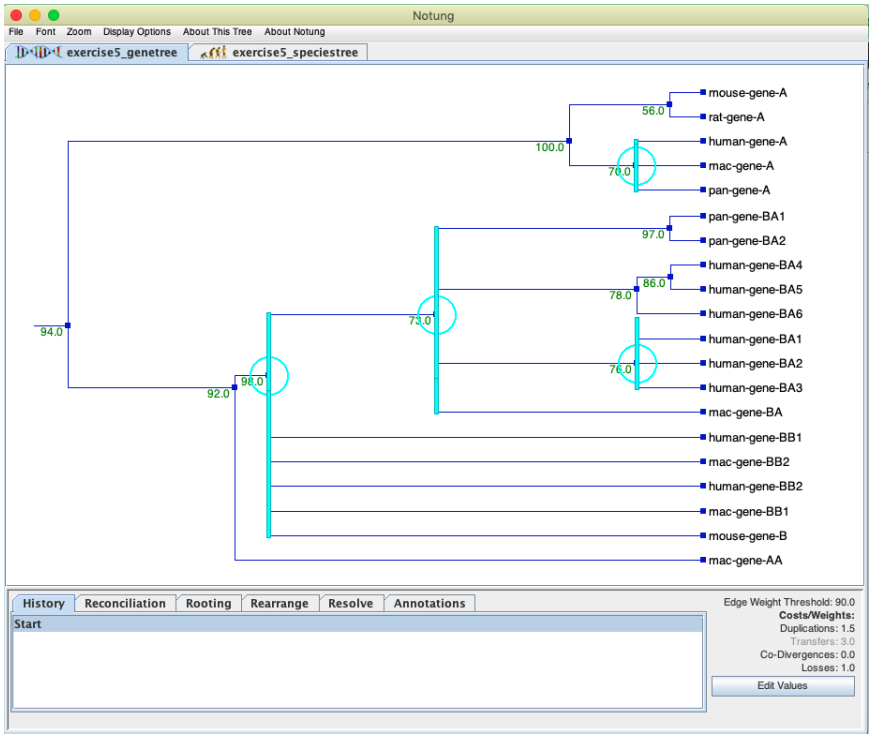


Figure E.12: The gene tree with polytomies highlighted.

Reconcile the gene tree with the species tree

1. Click the "**Reconciliation**" tab.

2. Click "**Reconcile/Rereconcile**."

3. In the "Reconciliation Options" dialog box, select `exercise5_speciestree` and "Prefix" and click "**Reconcile**."

   The reconciled tree appears in the tree panel. Note that it has a DL Score of 20.0, with ten duplications and five losses. Also note that some of the polytomies have more than one duplication associated with the node (ex: the polytomy with eight children has two duplications).

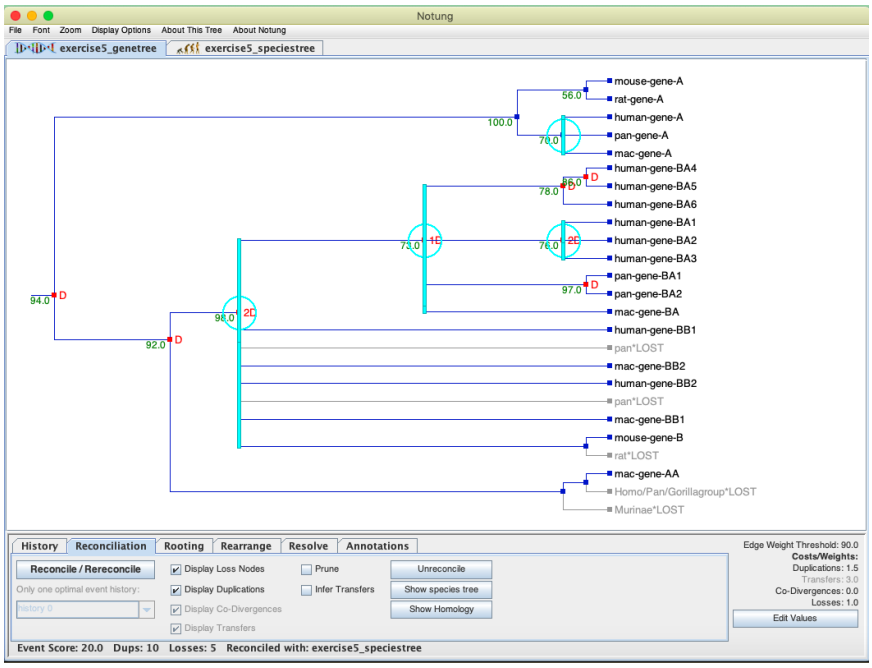


Figure E.13: The gene tree should now look like this.

Get general tree statistics for the gene tree

In this step you will gather some general statistics about the reconciled gene tree and the species tree.

1. Click on "**About This Tree → General Tree Statistics**"

   The General Tree Statistics window appears. In this window is information on both the gene tree, the reconciled gene tree, and the species tree. You may have to scroll down to view all the information.

   The General Tree Statistics Window should look like this.

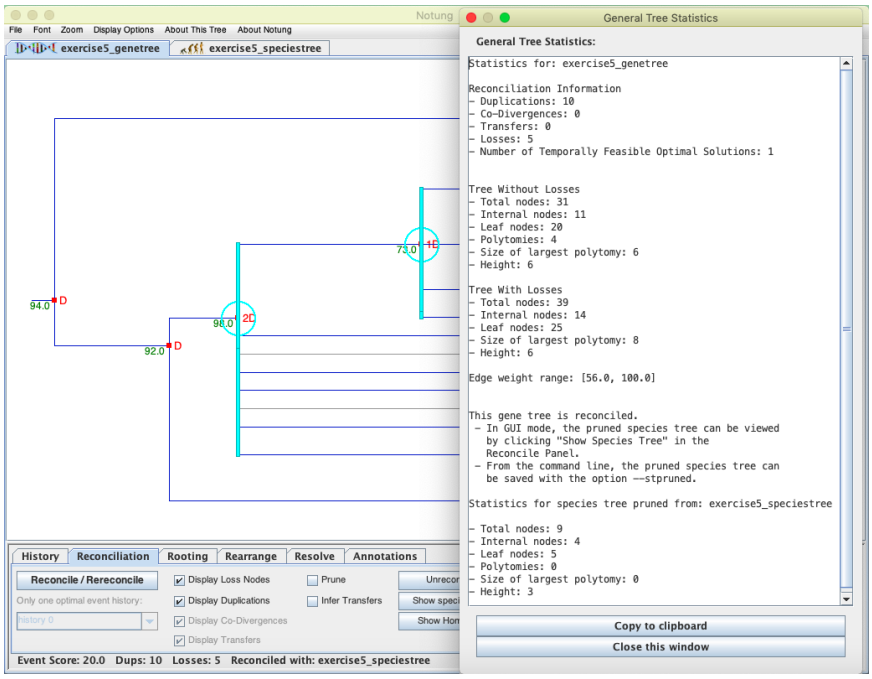


Figure E.14: The General Tree Statistics Window should look like this.

2. After reviewing this information, click "**Close this window**" to close the window and continue.

   For more information on the data in the General Tree Statistics window, see Chapter 3.4 - General Tree Statistics on page 20.

Resolve the polytomies in the gene tree

In this step, you will resolve all the polytomies in the gene tree, thus creating a binary gene tree.

1. Click the "**Resolve**" tab.

   The Resolve task panel opens below.

2. Make sure that the "Highlight polytomies" checkbox is selected.

   The polytomies in the gene tree are circled and highlighted in cyan.

3. Click "**Resolve Polytomies**."

   The resolved tree appears in the tree panel. Edges associated with the resolved polytomies are now colored cyan. This is the same tree as before, only now the polytomies have been resolved. The number of duplications and losses are identical to the reconciled tree, and even the duplication bounds are the same.

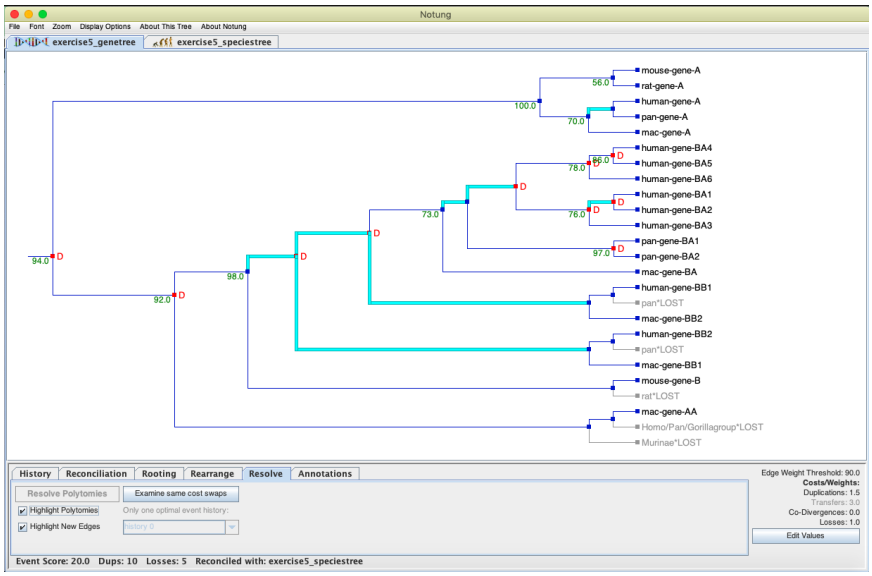


Figure E.15: The gene tree should now look like this.

Change the parameter values and view alternate event histories

In the previous steps, we reconciled and resolved the tree using the default parameter values (CD=1.5 and CL=1.0). For the default values, there is only one minimum cost tree. We now explore what happens when we reconcile the tree when duplications and losses are weighted equally.

1. Click the "**History**" tab.

   We must go back in the history before we change parameter values, as the tree has already been resolved and the change in values might effect the current resolution of the tree.

2. Go back to the pre-resolved step in the history by clicking "**Reconciled with exercise5_speciestree**."

   The tree panel shows the state of the tree before the polytomies were resolved.

3. Click the "**Edit Values**" button in the bottom-right corner of the program.

4. In the dialog box, change the Duplication Cost to 1.0.

5. Click "**Apply Changes**."

   Duplications and losses are now weighted equally, and the gene tree is automatically re-reconciled with the new parameter values.

6. Click the "**Reconciliation**" tab.

   The reconciled tree appears in the tree panel. There is now more than one alternate gene tree with the minimal DL Score. You are currently viewing `history 0`.

7. Click on the drop-down menu labeled "**Select an optimal event history**."

8. Select `history 1`.

   A different tree appears. This tree has a DL Score of 15.0, with ten duplications and five losses. This tree has the same duplications and losses as the tree reconciled with a duplication cost of 1.5 and a loss cost of 1.0 (see Figure E.14).

9. Click on the drop-down menu labeled "**Select an optimal event history**" and select `history 0`.

   A different tree appears. This tree also has a DL Score of 15.0, but has eleven duplications and four losses rather than the ten duplications and five losses in `history 1`. The large polytomy with seven children now has three duplications and one loss, whereas in `history 1` it had two duplications and two losses.

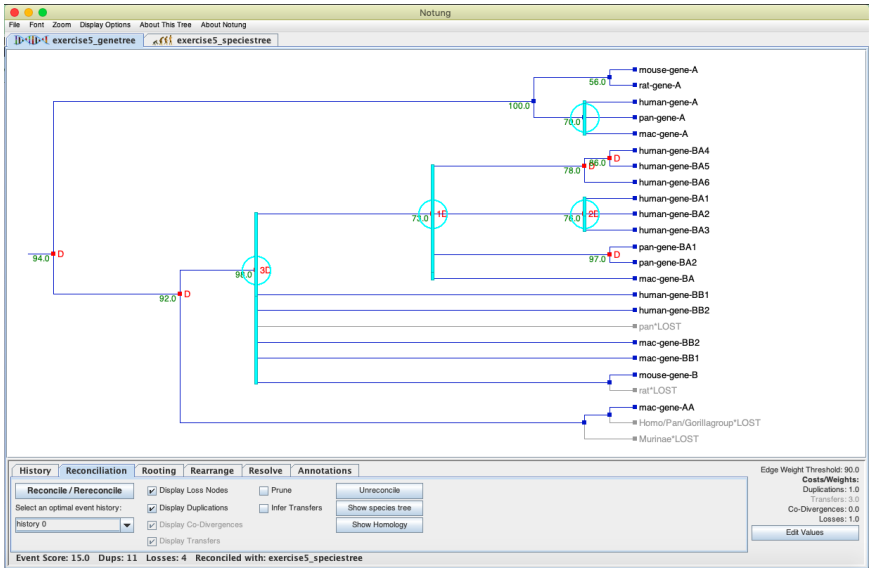


Figure E.16: The gene tree should now look like this.

Run the Rooting Analysis

1. Click the "**Rooting**" tab.

2. Click "**Run Rooting Analysis**."

   Many edges and one polytomy are colored red, which indicates that all of these components of the tree have the lowest root score.

Notice that the large polytomy is circled in red. Placing a root at a polytomy indicates that at least one edge in the binary resolution of the polytomy has the lowest root score.

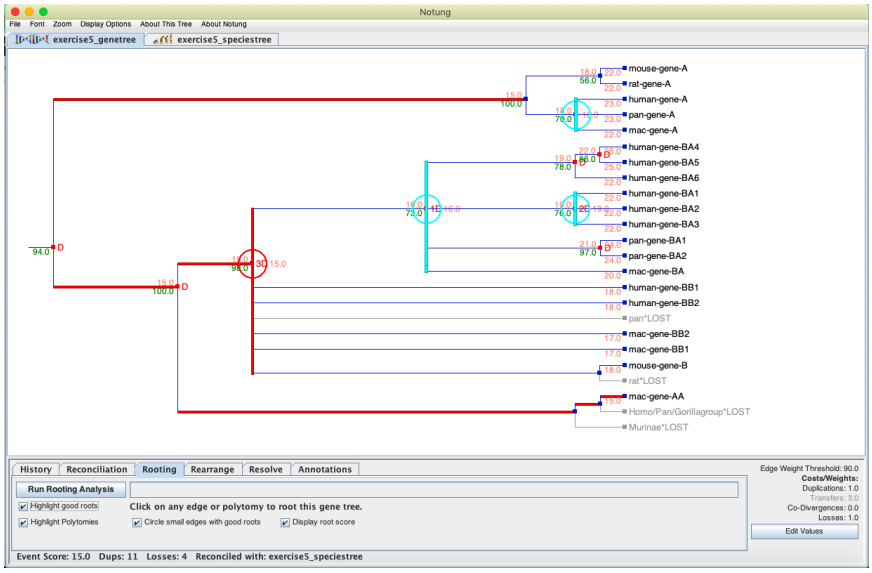


Figure E.17: The gene tree should now look like this.

3. Optional: Click the "**Display root score**" checkbox.

   Each edge and polytomy is labeled with its root score.

4. Select an optimal root by clicking on the polytomy with the red circle.

   The tree is rooted on the polytomy and the DL Score of the tree is still 15.0, with eleven duplications and four losses.

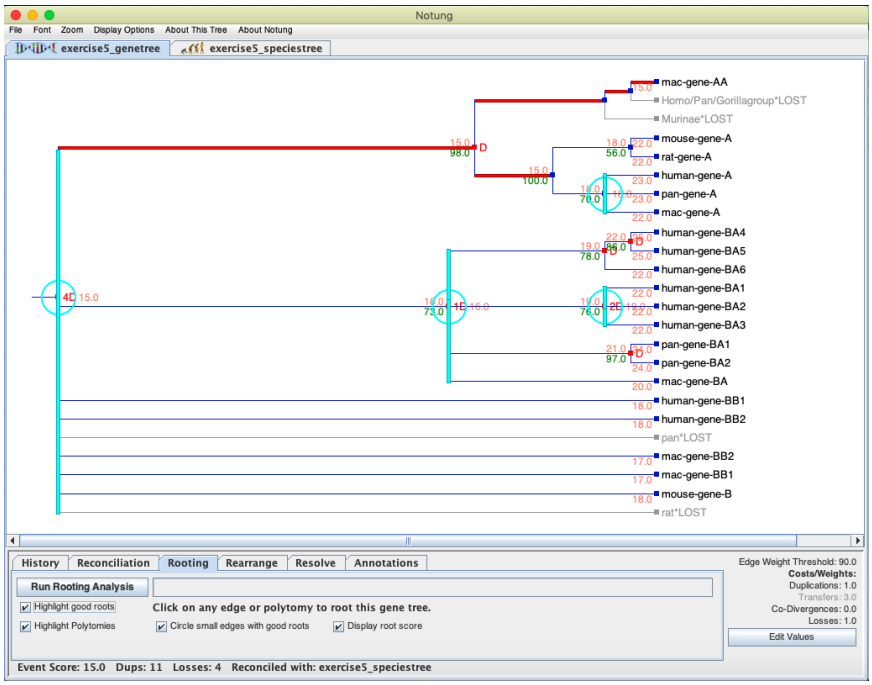


Figure E.18: The gene tree should now look like this.

Resolve the polytomies in the gene tree

In this step, you will resolve all the polytomies in the gene tree, thus creating a binary gene tree.

1. Click the "**Resolve**" tab.

   The Resolve task panel opens below.

2. Make sure that the "**Highlight polytomies**" checkbox is selected.

   The polytomies in the gene tree are circled and highlighted in cyan.

3. Click "**Resolve Polytomies**."

   The resolved tree appears in the tree panel. Edges associated with the resolved polytomies are now colored cyan.

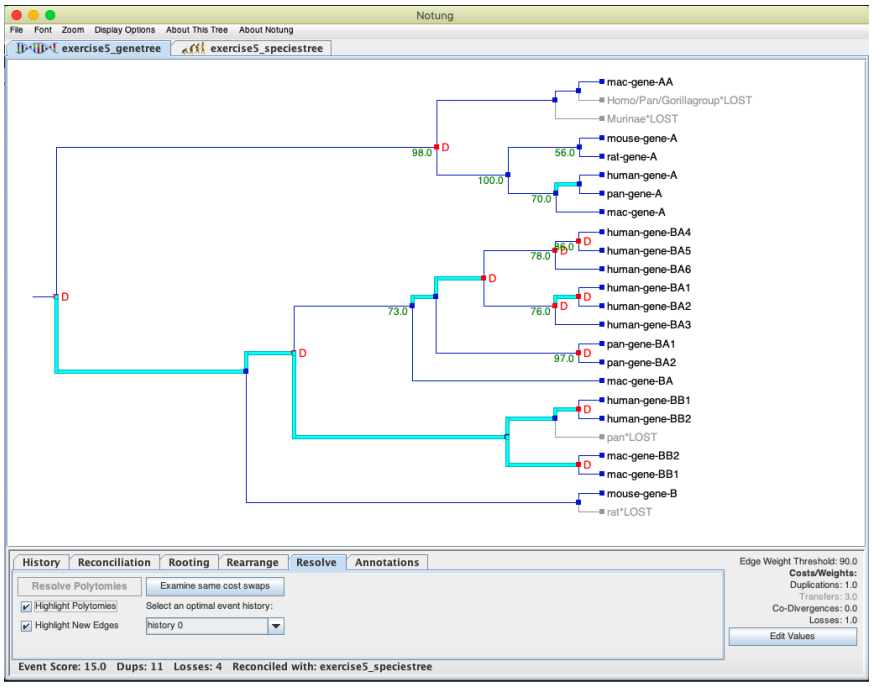


Figure E.19: The gene tree should now look like this.

View a different alternate event history

With these parameter values, there is more than one alternate gene tree with minimal DL Score. You are currently viewing `history 0`.

1. Click on the drop-down menu labeled "Select an optimal event history."

   This displays a list of available alternate event histories. You should see `history 0` and `history 1`.

2. Select `history 1`.

   A different tree appears. This tree also has a DL Score of 15.0, but has ten duplications and five losses instead of eleven duplications and four losses.

   Note that these alternate histories correspond to the same alternate histories that were presented after reconciliation.

Swap nodes in the resolved tree

Note that this tree groups `human-gene-BB1` with `mac-gene-BB2` and `human-gene-BB2` with `mac-geneBB1`. However, the tree that groups `human-gene-BB1` with `mac-geneBB1` and `human-gene-BB2` with `mac-gene-BB2` has the same score.

1. Click the "**Examine same cost swaps**" button.

   Nodes that can be interchanged without changing the DL Score or history implied by the polytomies are marked with enlarged light blue boxes.

2. Click the node for `mac-gene-BB1`.

   The node is now marked with a light blue triangle. Each node it can be swapped with is marked with a pink triangle. In this case, there is just one: the node leading to `mac-gene-BB2`.

3. Click the node for `mac-gene-BB2`.

   The nodes `mac-gene-BB1` and `mac-gene-BB2` are swapped. Once they have been swapped, they are temporarily highlighted with yellow triangles, so that you can see the results of the most recent action. Note that the **BB1** genes are now grouped together, and the **BB2** genes are together in the same subtree.

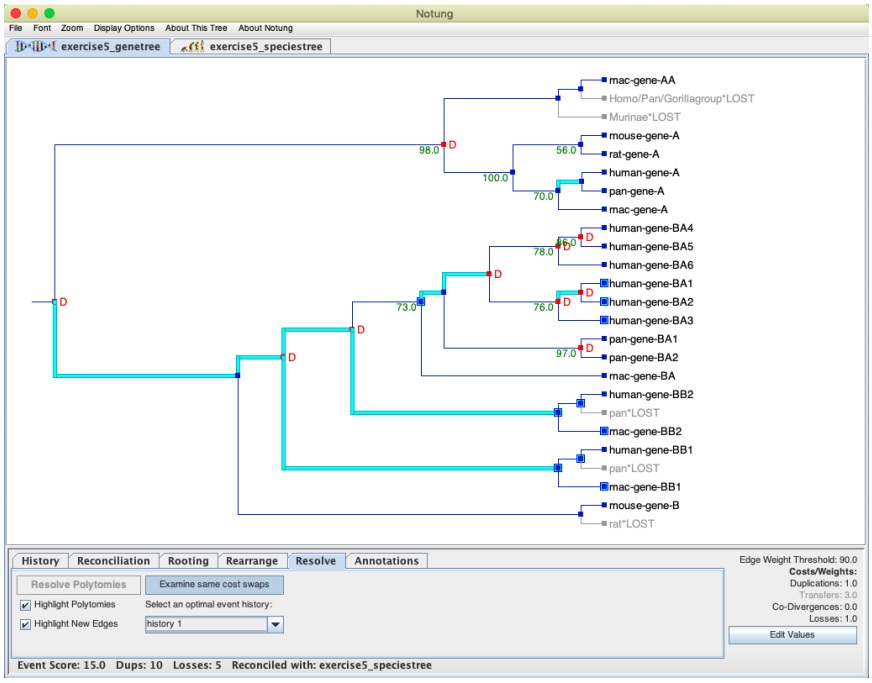


Figure E.20: The gene tree should now look like this.

Annotate the Gene Tree

This step will introduce you to Notung's annotations capabilities.

1. Click the "**Annotations**" tab.

   The Annotations task panel is displayed.

2. Click on the "**New**" button to add a new annotation.

   A box will appear to edit the new annotation.

3. In the space labeled "Please enter a title for the annotation", type in "-A" Select a color from the palate and click "**OK**".

   This will automatically annotate all the leaves that contain the string "-A" with the color you selected.

4. Click on the "**New**" button. In the space labeled "Please enter a title for the annotation," type in "-BA" and select a color from the palate and click "**OK**."

   This will automatically annotate all the leaves that contain the string "-BA" with the color you selected.

5. Click on the "**New**" button. In the space labeled 'Please enter a title for the annotation," type in "BB1" and select a color from the palate and click "**OK**".

   This will automatically annotate all the leaves that contain the string "BB1" with the color you selected.

6. Click on the "**New**" button. In the space labeled "Please enter a title for the annotation," type in "BB2" and select a different color from the palate and click "**OK**".

   This will automatically annotate all the leaves that contain the string "BB2" with the color you selected.

7. Click on the "**New**" button. In the space labeled "Please enter a title for the annotation," type in "BA 4, 5, 6" and select a different color from the palate and select the button labeled "I want to manually select the nodes and subgroups to add." Click "**OK**".

   This option lets you select the nodes to add to the annotation without searching for a substring.

8. Click on the node leading to the subtree with genes BA4, 5, and 6 in humans.

   Notice that these leaves were previously in the color selected in step 3. The leaves are a new color now because the newer annotation takes precedence.

9. Click on the "**New**" button. In the space labeled "Please enter a title for the annotation," type in "BA 1, 2, 3" and select a different color from the palate and select the button labeled "I want to manually select the nodes and subgroups to add." Click "**OK**."

10. Click on the node leading to the subtree with genes BA1, 2, and 3 in humans.

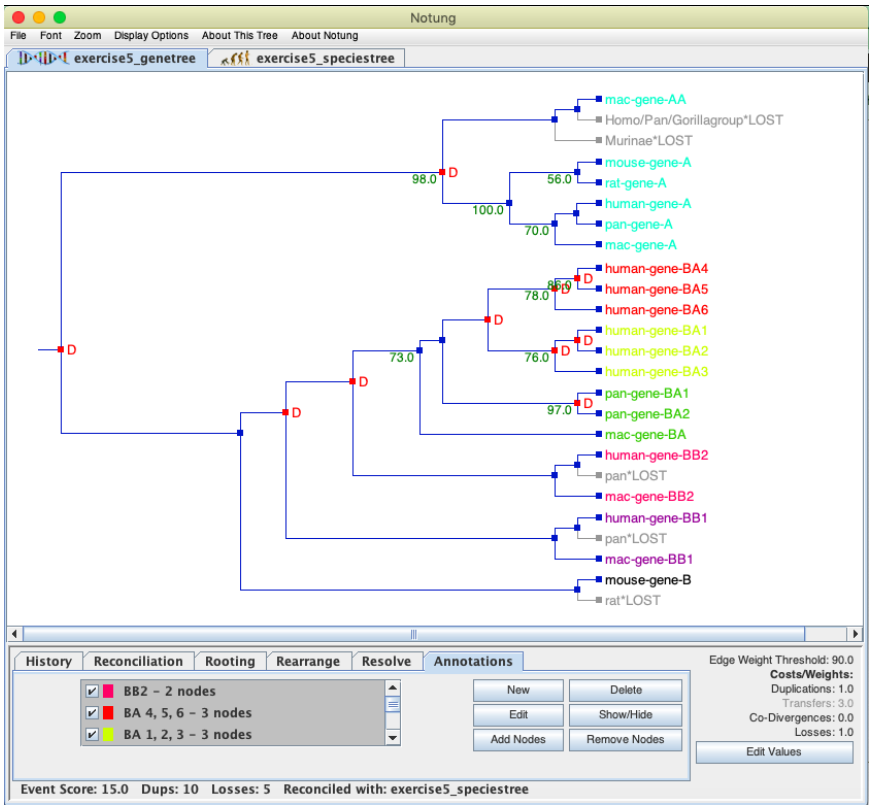


Figure E.21: The gene tree should now look something like this.

Rearrange the resolved tree

In this step, you will rearrange the gene tree to obtain the minimal DL Score. In this exercise, you have resolved the polytomies in the gene tree before rearranging the weak areas of the tree. However, it is possible to do both task at the same time while in the rearrangement mode. Both Resolve and Rearrangement are available because these two functions have different purposes. If you want to obtain a hypothesis of the binary gene tree, but wish to retain all the information in the gene tree, use the Resolve task mode. However, if you wish to consider edges with an edge weight below a certain value as uninformative, use the Rearrangement task mode.

1. Click the "**Rearrange**" tab.

2. Click the "**Highlight weak edges**" checkbox.

   Several edges in the reconciled tree are highlighted in yellow. These are edges with weights below the Edge Weight Threshold and are considered "weak." Weak edges may be rearranged to reduce the number of duplications and losses in the tree. Edges with weights above the threshold will not be rearranged.

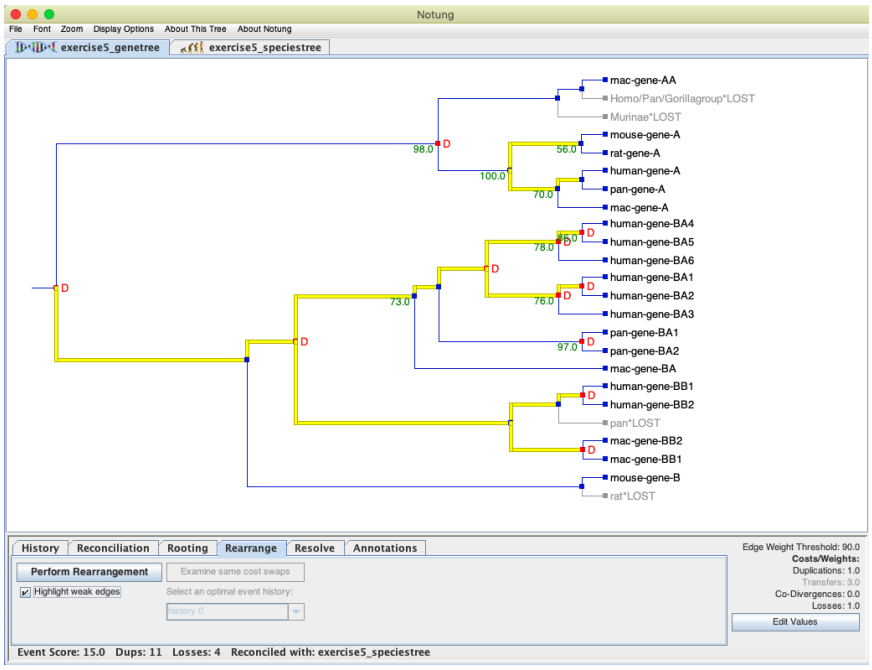


Figure E.22: The gene tree with weak edges highlighted.

3. Click "**Perform Rearrangement**."

The rearranged tree appears in the tree panel. It has a DL Score of 15.0, with twelve duplications and only three losses. Note that the score did not change; the rearranged tree is not necessarily "better" than the original tree.

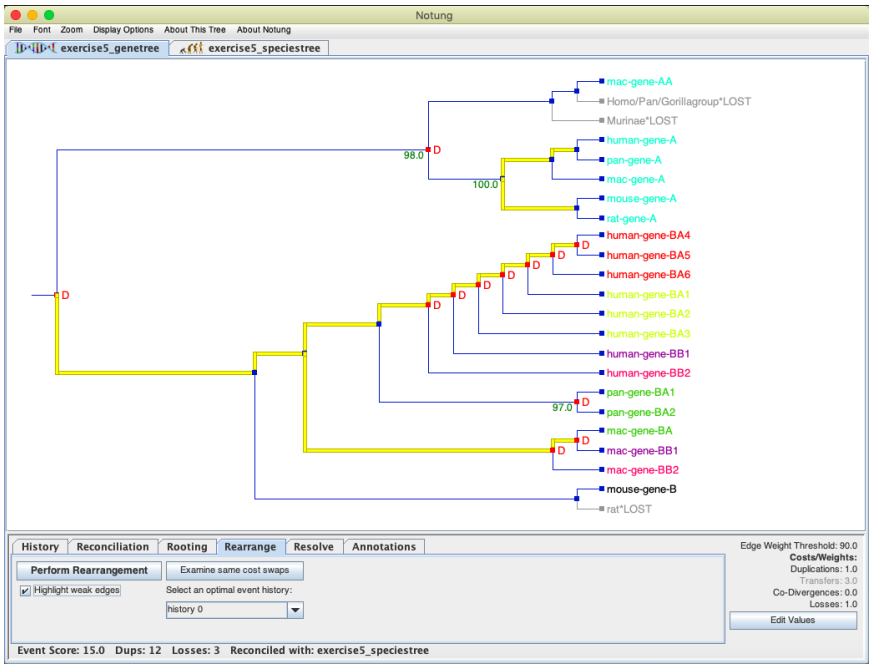


Figure E.23: The gene tree should now look like this.

View a different alternate event history

You are currently viewing `history 0`.

1. Click on the drop-down menu labeled "Select an optimal event history."

   This opens a list of available alternate event histories. You should see `history 0`, `history 1`, and `history 2`.

2. Select another history and examine the same cost swaps by clicking the "**Examine same cost swaps**" button.

   Nodes that can be interchanged without changing the DL Score are marked with enlarged light blue boxes. Try performing additional swaps to see how many alternate, minimum cost trees you can find.

3. See if you can find the original tree by changing the histories and examining same cost swaps.

   HINT 1: Select the history with ten duplications and five losses.

   HINT 2: Swap the subtree of BA1 and BA2 genes in "pan" with the LOST "pan" gene in the BA subtree.

   HINT 3: Swap the subtree of BA4, BA5, and BA6 in human with the node for BA3 in human.

E.6 Exercise 6 - Inferring transfers in a binary gene tree with a non-binary species tree

In this exercise, you will use Notung to infer transfers with duplications, losses, and ILS in the gene tree, genetree-bacteria with the non-binary species tree `speciestree-bacteria`.

Open the tree files

1. Click "**File → Open Gene Tree**" and open `genetree-bacteria`.

   This is an artificial tree made up for this exercise. Notice that this gene tree is non-binary and contains multiple polytomies.

2. Click "**File → Open Species Tree**" and open `speciestree-bacteria`.

Reconcile the gene tree with the species tree

1. Select `genetree-bacteria`.

2. Click "**Reconciliation**" tab.

3. Click "**Reconcile/Rereconcile**."

4. In the "Reconciliation Options" dialog box, select `speciestree-bacteria` and "Prefix" and click "**Reconcile**."

   This tree has been reconciled with the duplication-loss algorithm. Note that it has a DL score of 23with 4 duplications, 17 losses, and 1 co-divergence.

5. Check the "**Infer Transfers**" box.

   Now when you reconcile the tree you will include transfer inference.

6. Click "**Reconcile/Rereconcile**."

7. A message will appear saying "No temporal feasible solutions found."

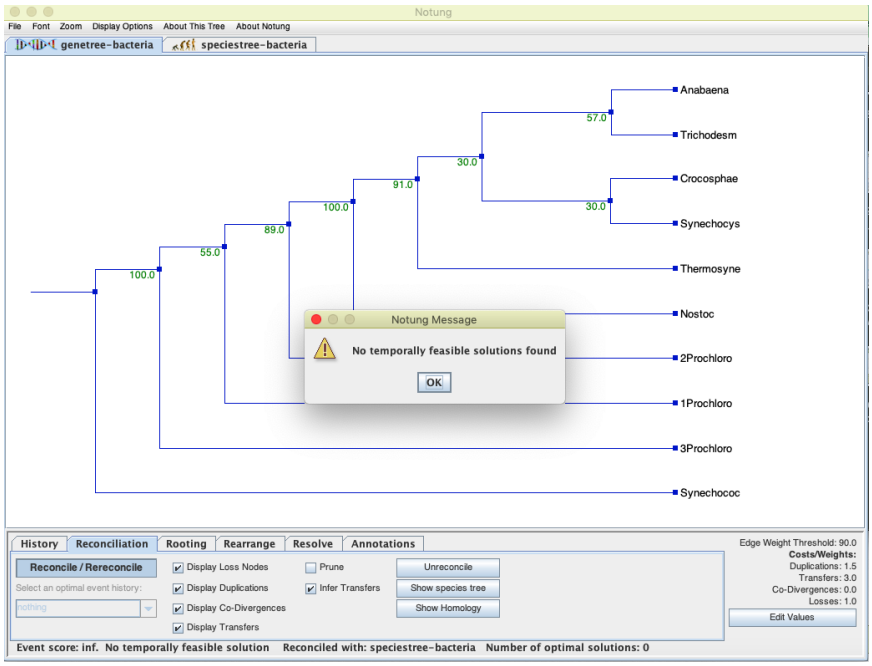


Figure E.24: A warning will pop up when there's no temporal feasible solution

Click "**OK**." This means that with these costs and this root, all optimal solutions were temporally infeasible. For an explanation of this error, see Chapter 5.2.1 - Temporal Feasibility in Notung on page 43.

Run the Rooting Analysis

Let's see if another root will produce a temporally feasible solution.

1. Click "**Rooting**."

2. Click "**Run Rooting Analysis**."

3. In the "Reconciliation Options" dialog box, select `speciestree-bacteria` and "Prefix" and click "**Reconcile**."

4. Click the "**Display root score**" checkbox.

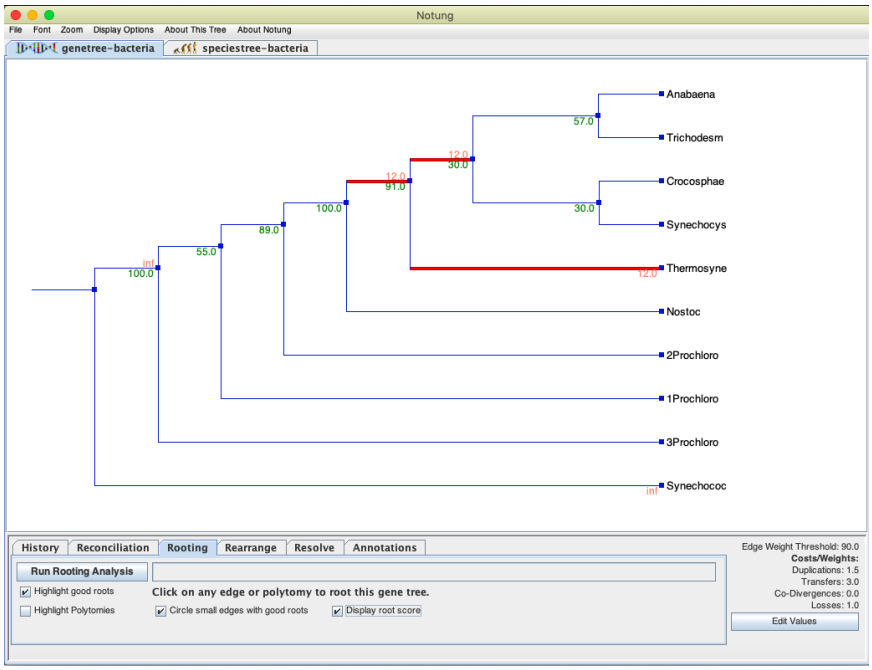


Figure E.25: Root scores will be displayed on the gene tree branches

Optimal edges that have been checked for temporal feasibility will be labeled with a roots score. Notice the current root edge has a score of inf. This means there are no temporally feasible solutions for this root, and the root score is infinity. see Chapter 6 - Rooting Mode on page 68 for more information. The edge(s) with the lowest score and at least one feasible solution are highlighted in red.

5. Notung is suggesting that the gene tree be rerooted on the red branches. Click on the red branch that is closer to the root (left most).

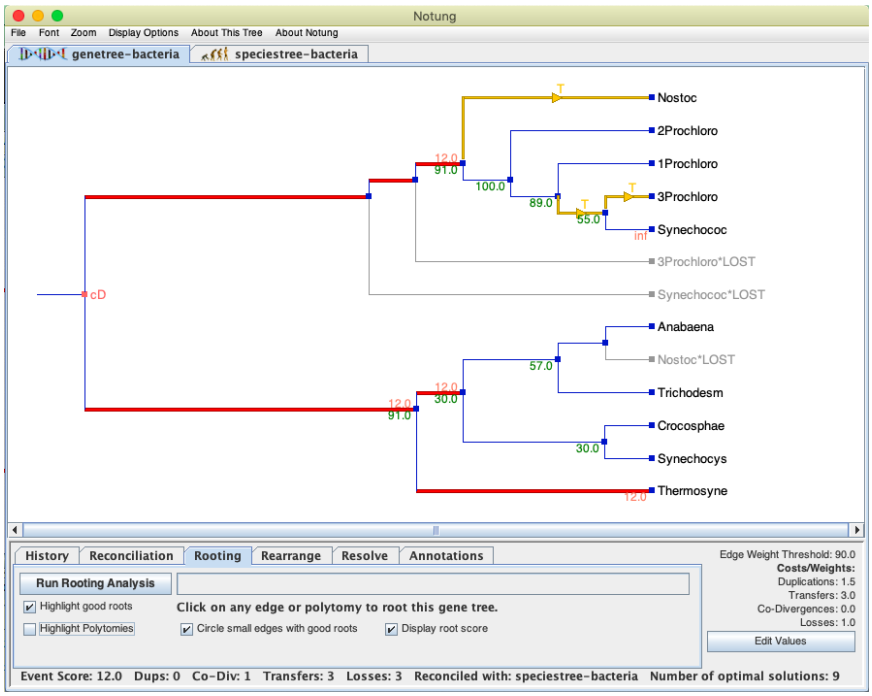


Figure E.26: What the gene tree will be once rooted at one of the red branches

The tree is now rooted and has a DTL score of 12with 0 duplications, 3 transfers, 3 losses and 1 co-divergence.

View alternate event histories

1. Click on the "**Reconciliation**" tab.

   Notice that some nodes are circled and colored green. This indicates areas that will be affected by alternative optimal histories. see Chapter 5 - Reconciliation Mode on page 40 for more information.

2. Click on any circled node.

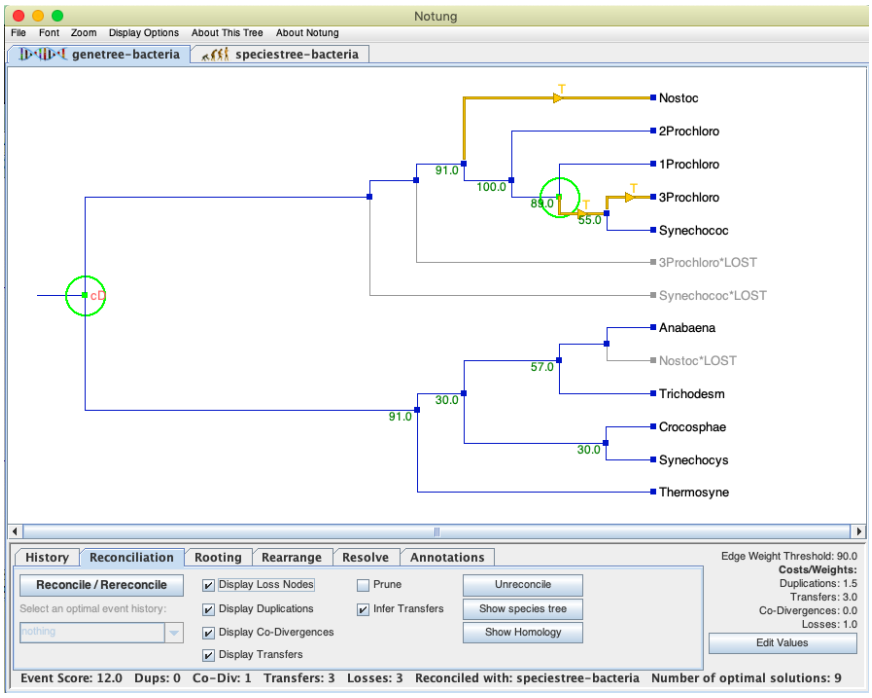


Figure E.27: Green circles on the nodes indicate alternative event history, and are clickable

This will cycle through all other histories below that node. Clicking on the circled node(s) closest to the root will cycle through all the histories. The total number of alternative optimal solutions is indicated in the status bar.

Get event details for the history

1. Click on "**About This Tree → Event Summary**" menu item.

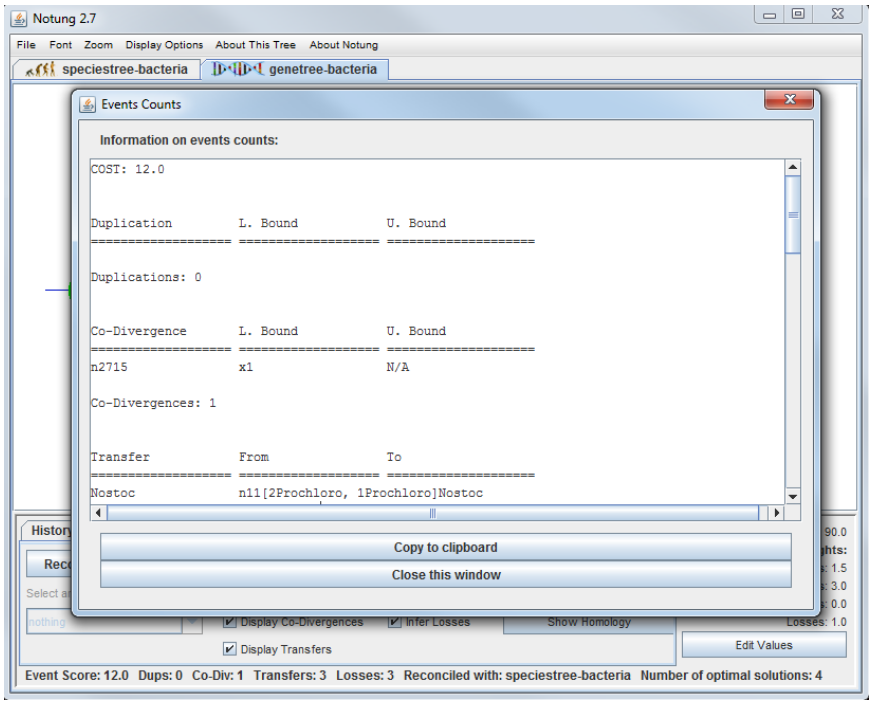


Figure E.28: Events Summary window

The "Event Summary" window appears. You will see the DTL score for the current history. Below that, you will see a table listing gene duplication events. The third table lists all the transfers that Notung inferred. The first column lists the name of the ancestral gene which was transferred. The "From" column lists the donor species that is the source of the transferred gene, while the "To" column lists the name of the ancestral recipient species. The next table lists the number of times each species lost a gene. The final table provides a summary of all events involving each species.

2. Close the window by clicking the "**Close this window**" button.

Appendix F

Troubleshooting

Problem	Possible Causes	Solutions
When I tried to reconcile the trees, I received this error message: "None of the species labels in this tree can be found in the species tree. Try checking your reconciliation settings."	• The species labels in the gene tree leaf node names are not compatible with the species labels in the species tree. • The "Specify Species Label" setting in the Reconciliation Options dialog box has been set incorrectly. • The incorrect species tree has been selected for reconciliation.	• Check the species labels in the gene tree to make sure they match the species labels in the species tree. • In the Reconciliation Options dialog, make sure you select the appropriate naming convention for species labels. • In the Reconciliation Options dialog, make sure you select the appropriate species tree for reconciliation.
The edge weights on the gene tree are not what you expected.	Notung has mistaken the branch length values in the Newick file for edge weight values. See Appendix A.6 - Location of Edge Weight Values on page 131	First, open the gene tree file in a text editor to determine the location of edge weight values. Then, click "**Display Options → Select Location of Edge Weights**" and set the location of Edge Weights appropriately.

Problem	**Possible Causes**	**Solutions**
My gene tree should have edge weights, but when I load the tree, weights are not displayed on some branches.	The gene tree file is supposed to be in Newick, NHX or Notung format, but contains a typo or formatting error, affecting the edge weight location.	Open the original tree file in a tree editing program or text editor and correct any formatting errors. *NOTE*: Some formats are case-sensitive.
When I tried to reconcile the gene tree with the species tree, I received this message: "There are no species trees to reconcile with." -or- My species tree is not listed in the drop down menu in the Reconciliation Options dialog box.	You have opened a species tree as a gene tree.	Reopen the desired species tree as a species tree using "**File → Open Species Tree**" or "**Ctrl-Shift-O**".
After reconciliation, I found lost genes in unrecognizable species, such as "n101."	The gene was lost in an ancestral species that was not given a label in the original species tree file. When internal node names are not specified in the input file, Notung generates them using an arbitrary counting system (ex: n101).	Use "**Display Options → Display Internal Node Names**" to examine internal species names in the species tree. If you prefer taxonomic names, use a tree editing program or text editor to add real species names to internal nodes in the species tree.
When I tried to open a tree, I received this message: "An error occurred while opening your file. Please check the format." Or "An error occurred while opening your file. Node had malformed information."	• The gene tree file is supposed to be in Newick, NHX or Notung format, but contains a typo or formatting error. • The gene tree file is in a format Notung does not accept, (ex: Nexus).	• Open the original tree file in a tree editing program or text editor and correct any formatting errors. • Convert the file to Newick or NHX file format. See Appendix A - File Formats on page 126 for more information about file formats.

Problem	**Possible Causes**	**Solutions**
Notung reports that you do not have a recent enough version of Java, but you have the latest version installed.	You have multiple versions of Java installed.	• On Windows, bring up the properties window for the Notung-2.7 jar file. Check the "Opens With" field - if the wrong version of java is listed, change it so that the right version of java is being used. • On Linux or Mac, type `java -version` - this will tell you which version of Java is being used. If it is incorrect, alter your path environment variable to include the proper version of Java.
The species tree file I created using the NCBI Taxonomy Browser contained non-ASCII characters.	As part of its file construction, the NCBI Taxonomy Browser includes some non-ASCII characters.	These characters are ignored by Notung, but you can open the tree file in a text editor and delete the non-ASCII characters.
The species tree file I created using the NCBI Taxonomy Browser contained 4's.	As part of its file construction, the NCBI Taxonomy Browser includes a branch length of 4 for every edge in the species trees it produces.	These branch lengths are ignored by Notung, but you can open the tree file in a tree editing program or text editor and delete the branch lengths.

Problem	**Possible Causes**	**Solutions**
The names of internal nodes in my gene tree change over time.	• The gene tree file does not specify internal node names and has been reloaded. When internal node names are not specified in the input file, Notung generates them using an arbitrary counting system (ex: n101). • Node names were given in the original tree file, but additional nodes have been added, because either rearrangement or resolve has been performed. Added nodes are assigned names that begin with an 'r' and are followed by numbers (ex: r245).	• If you want the internal node names to be the same every time the tree is opened, use a tree editing program or text editor to add names to internal nodes in the gene tree. • Notung cannot track internal nodes that are temporary or not present in the original file. If you need permanent names for these nodes, save the file and use a tree editing program or text editor to specify names for these nodes.
I use the <Tab> key to navigate to a different button in a pop-up box, but when I hit the <Enter> key, the selected button is not engaged.	This is a problem with some versions of Java. The <Tab> key option to navigate to different buttons does not select the "highlighted" button. When the <Enter> key is pressed, the originally selected button is used.	Use the mouse to select buttons in the windows
I have added a node to an annotation, but the node does not appear in the correct color.	There are conflicting annotations - the node corresponds to more than one annotation and is currently being described by another annotation. Annotations have precedence - those annotations added later will always take precedence over earlier annotations.	Manually remove the node from the other annotations, or check the other annotations and remove any search strings that identify the node of interest. See Chapter 10 - Annotations on page 87 for more information.

Bibliography

[1] J. Andersson. Horizontal gene transfer between microbial eukaryotes. *Methods Mol Biol*, 532:473–487, 2009.

[2] K. Chen, D. Durand, and M. Farach-Colton. Notung: A program for dating gene duplications and optimizing gene family trees. *J Comput Biol*, 7(3/4):429–447, 2000.

[3] R. R. Copley, I. Letunic, and P. Bork. Genome and protein evolution in eukaryotes. *Curr. Opin. Chem. Biol.*, 6(1):39–45, Feb 2002.

[4] Charlotte A. Darby, Maureen Stolzer, Patrick J. Ropp, Daniel Barker, and Dannie Durand. Xenolog classification. *Bioinformatics*, 33(5):640–649, Mar 2017.

[5] J. Degnan and N. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, 24:332–340, Jun 2009.

[6] J. Doyon, V. Ranwez, V. Daubin, and V. Berry. Models, algorithms and programs for phylogeny reconciliation. *Brief Bioinform*, 12:392–400, Sep 2011.

[7] D. Durand, B. Halldorsson, and B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, 13(2):320–335, 2006.

[8] I. Ebersberger, P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, et al. Mapping human genetic ancestry. *Mol Biol Evol*, 24:2266–2276, Oct 2007.

[9] S. Edwards. Is a new and general theory of molecular systematics emerging? *Evolution*, 63:1–19, Jan 2009.

[10] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2003.

[11] J. Felsenstein. *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington, Seattle, 2005.

[12] W. M. Fitch. Homology: a personal view on some of the problems. *Trends Genet*, 16(5):227–231, May 2000.

[13] M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *RECOMB 2004: Proceedings of the Eigth International Conference on Research in Computational Biology*, ACM Press, pages 347–356, New York, NY, USA, 2004. ACM.

[14] R. Hudson. Gene genealogies and the coalescent process. In *Oxford surveys in evolutionary biology*, volume 7, pages 1–44. Oxford University Press, 1990.

[15] D. H. Huson and C. Scornavacca. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol*, 3:23–35, Nov 2011.

[16] H. Lai, M. Stolzer, and D. Durand. Fast heuristics for resolving weakly supported branches using duplication, transfers, and losses. In *RECOMB International Workshop on Comparative Genomics*, volume 10562, pages 298–320. Springer, Springer, 2017.

[17] W. Maddison. Reconstructing character evolution on polytomous cladograms. *Cladistics*, 5:365–377, 1989.

[18] W. Maddison. Gene trees in species trees. *Syst. Biol.*, 46(3):523–536, 1997.

[19] L. Nakhleh. Evolutionary phylogenetic networks: models and issues. In L Heath and N Ramakrishnan, editors, *The Problem Solving Handbook for Computational Biology and Bioinformatics*, pages 125–158. Springer, 2010.

[20] L. Nakhleh, D. Ruths, and H. Innan. Gene trees, species trees, and species networks. In R. Guerra and D. Goldstein, editors, *Meta-analysis and Combining Information in Genetics and Genomics*, pages 275–293. CRC Press, 2009.

[21] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A phylogenetic approach*. Blackwell Science, Malden, MA, 1998.

[22] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol Biol Evol*, 5(5):568–583, 1988.

[23] D. Pollard, V. Iyer, A. Moses, and M. Eisen. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet*, 2(10):e173, Oct 2006.

[24] M.H. Serres, A.R.W. Kerr, T.J. McCormack, and M. Riley. Evolution by leaps: gene duplication in bacteria. *Biol Direct*, 4:46, Nov 2009.

[25] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers, and incomplete lineage sorting with non-binary species trees. *Bioinformatics*, 28(18):i409–i415, 2012.

[26] D. L. Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, MA, 2002.

[27] F. Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, Oct 1983.

[28] N. Takahata and M. Nei. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2):325–344, Jun 1985.

[29] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *TCBB*, 8:517–535, Mar/Apr 2011.

[30] B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J Comput Biol*, 15(8):981–1006, 2008.

[31] O. Zhaxybayeva and W. Doolittle. Lateral gene transfer. *Curr Biol*, 21:R242–R246, Apr 2011.

[32] C. M. Zmasek and S. R. Eddy. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17(4):383–4, Apr 2001.