

# **MALIN: Maximum Likelihood Analysis of Intron Evolution User's Guide**

Miklós Csűrös

Department of Computer Science and Operations Research  
Université de Montréal  
Montréal, Québec, Canada

April 10, 2008

## License

The MALIN software package is distributed under the terms of the BSD license, as shown below.

Copyright © 2008, Miklós Csűrös  
All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the *Université de Montréal* nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

# Contents

<b>1</b>	<b>Overview</b>	<b>4</b>
1.1	Introduction . . . . .	4
1.2	Availability . . . . .	5
1.3	Background . . . . .	5
1.4	Basic design concepts . . . . .	7
1.4.1	Sessions and the work area . . . . .	7
1.4.2	Browsers, primary items and views . . . . .	7
<b>2</b>	<b>Using MALIN</b>	<b>9</b>
2.1	Sessions . . . . .	9
2.2	Data . . . . .	10
2.3	Alignments . . . . .	11
2.3.1	Opening alignment files . . . . .	11
2.3.2	Alignment file format: organism tags . . . . .	11
2.3.3	Alignment file format: intron positions . . . . .	11
2.3.4	Alignment file format: an example . . . . .	12
2.3.5	Alignment panel . . . . .	12
2.3.6	Alignment panel: table . . . . .	13
2.3.7	Alignment panel: graphical alignment display . . . . .	13
2.3.8	Alignment panel: conservation criteria . . . . .	14
2.4	Intron table . . . . .	15
2.4.1	Intron table: ambiguous characters . . . . .	16
2.4.2	Intron table file format . . . . .	16
2.4.3	Intron table: table panel . . . . .	17
2.5	Rates . . . . .	18
2.5.1	Rate model . . . . .	18
2.5.2	Rates panel . . . . .	18
2.5.3	Rates panel: table . . . . .	19

2.5.4	Rates panel: graphical display of rate variation . . . . .	20
2.5.5	Rates panel: graphical display of branch rates . . . . .	20
2.5.6	Rate computation . . . . .	21
2.5.7	Rate optimization: general parameters . . . . .	21
2.5.8	Rate optimization: initial model . . . . .	22
2.5.9	Rate optimization: rate variation . . . . .	22
2.5.10	Rate optimization: advanced options . . . . .	23
2.5.11	Rate optimization: computing . . . . .	24
2.6	Analysis: shared presence . . . . .	25
2.7	Analysis: Dollo parsimony . . . . .	26
2.8	Analysis: posteriors . . . . .	28
2.9	Analysis: site histories . . . . .	29
2.10	Analysis: bootstrap . . . . .	31
<b>A</b>	<b>Test data</b>	<b>34</b>
<b>B</b>	<b>Command-line usage</b>	<b>36</b>
B.1	Introduction . . . . .	36
B.2	File types and formats . . . . .	36
B.3	Rate optimization . . . . .	37
B.4	Computation of posterior predictions . . . . .	39
B.5	Computation of Dollo parsimony . . . . .	39

# Chapter 1

## Overview

### 1.1 Introduction

MALIN is a software package for the analysis of eukaryotic gene structure evolution. It provides a graphical user interface for various tasks commonly used to infer the evolution of exon-intron structure in protein-coding orthologs. The implemented tasks include the following.

- Identification of homologous splice sites in annotated protein sequence alignments.
- Computation of primary statistics about introns in homologous sites (“shared introns”).
- Estimation of ancestral intron content, intron losses and gains by Dollo parsimony.
- Estimation of intron loss and gain rates in a probabilistic model.
- Estimation of ancestral intron content, intron losses and gains in a probabilistic model.
- Inference of evolutionary histories at individual sites.
- Error estimation for rates and histories by bootstrap.

## 1.2 Availability

MALIN is written entirely in Java, and, thus, can be used in different operating systems, including Mac OS X, Microsoft Windows, and various Unix/Linux versions. The software is packaged in a JAR file, and can be executed in Java versions 1.5 and above.

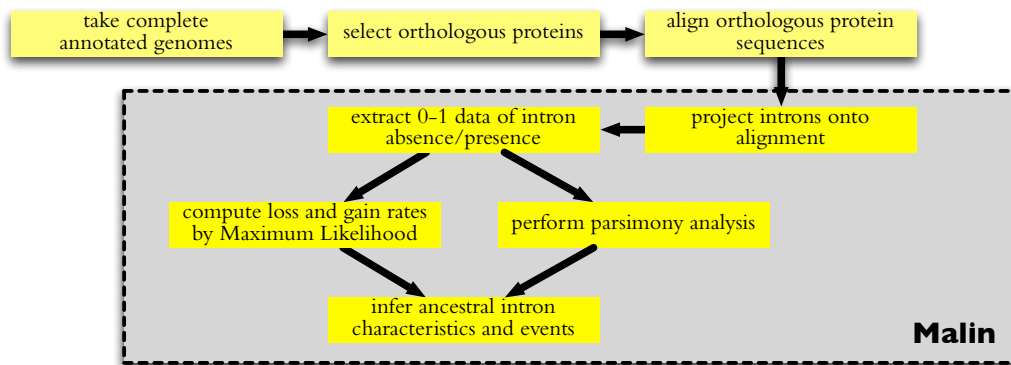
**Mac OS X** I have written the software using a Mac, and went to some extent to integrate the Java executable into a native-looking application. The JAR file is bundled as `Malin.app`, which you can just run directly by double-clicking on it.

**Microsoft Windows** You need to have a Java Virtual Machine on your computer in order to run MALIN. You could download, for instance, Sun's Java Runtime Environment from <http://www.java.com/>, which is the JRE I used in the testing. You will probably need to enable larger memory usage for the JVM than the default setting, which you get by double-clicking on the JAR file. You can launch MALIN via the provided MS-DOS batch file that sets the heap space for the JVM to 1000 Megabytes. Edit the batch file manually, if necessary.

**Unix/Linux** You can run MALIN from the command line, launching `java -jar Malin.jar`. You will probably need to enable larger memory usage for the JVM than the default setting, which you can do by launching MALIN as `java -Xmx1024M -jar Malin.jar`. The `-Xmx` option here sets the Java heap space to 1 Gigabytes: you can experiment with other settings appropriate for your computer and data set.

## 1.3 Background

Rogozin et al. (2005) give an overview of the analysis pipeline for eukaryotic gene structure evolution, which was employed in a number of recent studies starting with the investigations of Rogozin et al. (2003). MALIN can perform the tasks downstream of ortholog identification and alignment.



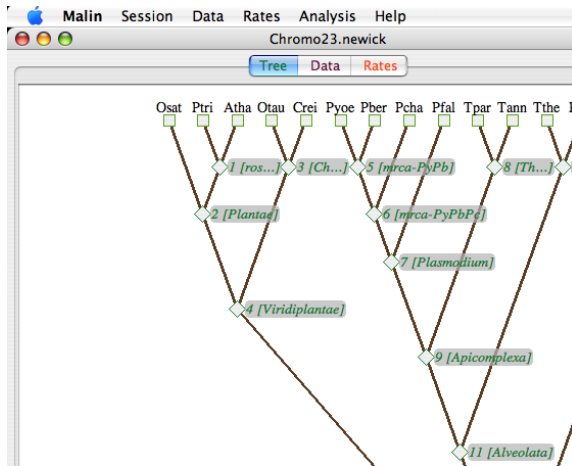
In order to infer if spliceosomal introns are in homologous positions, splice sites need to be projected onto coding sequences, and then homology is established from the protein alignments.



An *intron table* is constructed from the intron-annotated alignments. The table is a binary table of intron presence and absence in homologous sites across the studied organisms. The absence-presence patterns can be analyzed by Dollo parsimony (with the assumption that intron gains are rare events), or by probabilistic models of intron evolution. MALIN works with the likelihood framework that I have been developing (Csűrös 2005; Csűrös et al. 2007; Csűrös et al. 2008). The corresponding probabilistic model has branch-specific intron gain and loss rates, as well as rates-across-sites variation.

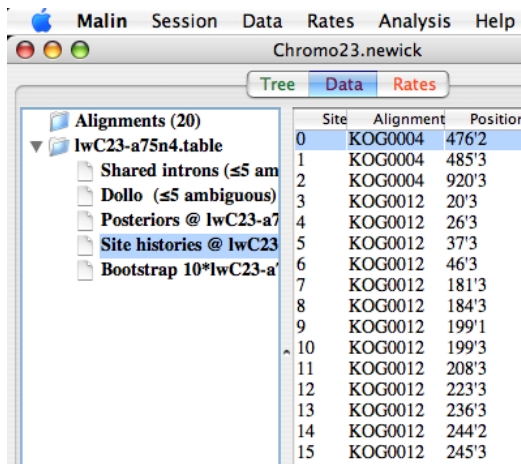
## 1.4 Basic design concepts

### 1.4.1 Sessions and the work area



MALIN operates with *sessions*: each session is associated with a fixed species phylogeny. More than one session may be open at one time: e.g., the same data set may be analyzed with different phylogenies simultaneously. A session has three main components, represented by the tabs of the displayed workspace: a species phylogeny (Tree), a *browser* for data sets and analysis results (Data) and a browser for probabilistic models (Rates).

### 1.4.2 Browsers, primary items and views



The Data and Rates tabs are attached to browser displays. A browser consists of a hierarchy on the left, and an information panel on the right, corresponding to the item selected in the hierarchy. Primary items in the hierarchy (depicted as folders of a file system) are alignments or data tables (under the Data tab), or rate models (under the Rates tab). Primary items may have *views*, which correspond to various analysis tasks. Views are descendant nodes in the hierarchy (depicted as documents or bullets, depending on the operating system).

Nodes of the hierarchy have small associated popup menus which you can bring up by right-clicking on them (or by Ctrl-click on a Mac). The popup menu items include the removal of the node from the browser, and possibly saving options.



Intron tables and rate models can be saved, and views are typically *exported* into text files. (The difference is that exported views cannot be loaded later, but saved tables and rate models can.)

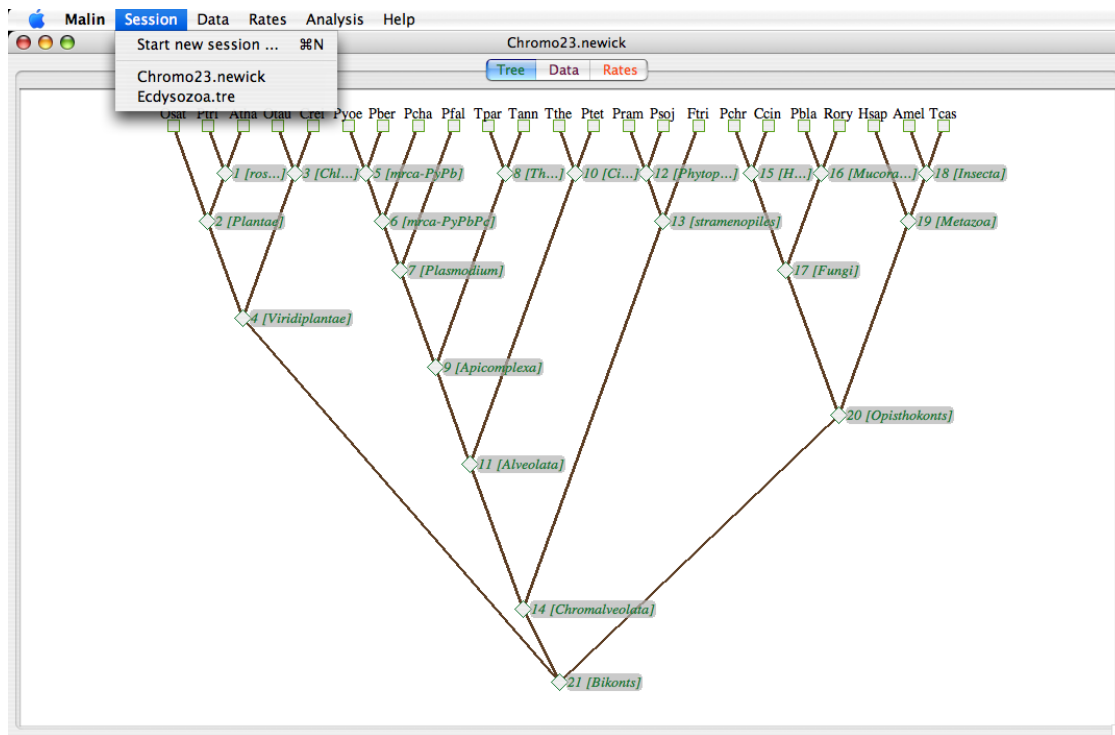
The browsers operate with split panes. You can resize the panes by dragging the dividers with the mouse, or even expand a pane completely by clicking on the little triangles on the bottom or the far right of the dividers.

# Chapter 2

## Using MALIN

### 2.1 Sessions

Data analysis in MALIN starts with opening a new session (Menu: Session → Start new session...). A session is opened by loading a species phylogeny. The phylogeny is expected to be in Newick format (<http://evolution.genetics.washington.edu/phylip/newicktree.html>) used by Phylip and other fine software packages for molecular evolution. The branch lengths of this phylogeny are ignored. The inner nodes of the tree may have more than two children: MALIN can deal with arbitrary multifurcations.



The phylogeny is displayed under the Tree tab. The inner nodes of the tree are numbered as 1,2,... (in a postorder traversal). It may be useful to name the inner nodes of the phylogeny, which is possible in Newick format:

```
((Pyoe, Pfal) "Plasmodium", (Tann, Tpar) "Theileria") "root";
```

For convenience in the graphical user interface, it is recommended that you use short names (3–4 letters) for the terminal taxa.

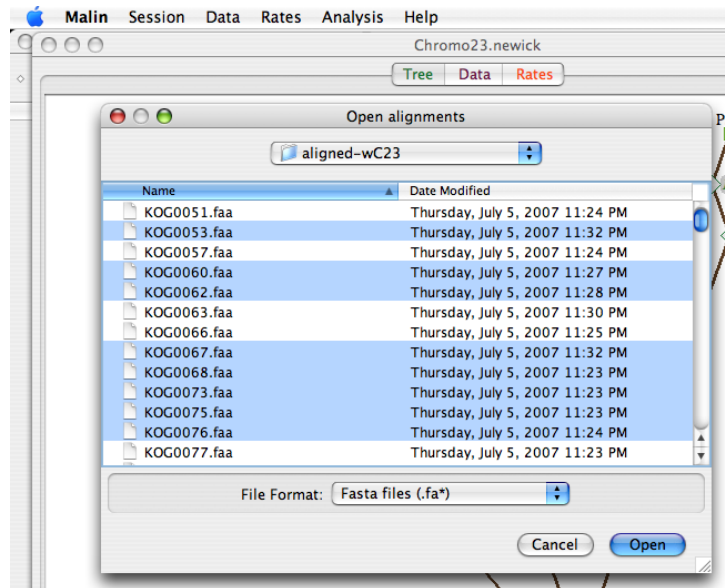
The open sessions are listed under the Session menu, where they can be selected to switch back and forth between them.

## 2.2 Data

Analysis tasks can be performed on intron tables, which are the binary tables of intron presence-absence. MALIN can construct such tables from annotated protein sequence alignments, or load precomputed tables directly.

## 2.3 Alignments

### 2.3.1 Opening alignment files



One or more Fasta files of protein alignments can be opened at a time (Menu: Data → Load alignments...). Each file is assumed to contain the alignment of a set of orthologous protein-coding genes.

Intron positions and organism identifiers are given in the Fasta sequence headers. Multiple files are selected in the usual manner for your operating system (e.g., shift+click for range selection).

### 2.3.2 Alignment file format: organism tags

The Fasta sequence headers need to specify the organism to which each sequence belongs. A sequence will be ignored if there is no terminal taxon for it in the session's phylogeny. Conversely, there will be a “missing ortholog” tag attached to each organism for which there is no sequence in an alignment. Organisms are specified by the /organism= tag in the header.

### 2.3.3 Alignment file format: intron positions

Intron positions are specified in the header line for each sequence. The program looks for a parenthesized list with the syntax {i  $p_1, p_2, \dots$  i}. Here  $p_1, p_2$ , etc. are intron positions with respect to the coding sequences (CDS):  $p_i = 1$  is the position after the first nucleotide position of the first codon (i.e., a phase-1 splice site within the first codon),  $p_i = 2$  is after the second nucleotide position of the

first codon,  $p_i = 3$  is the position between the first and the second codon, and so on. Notice that the positions are numbered with respect to the original, ungapped, CDS.

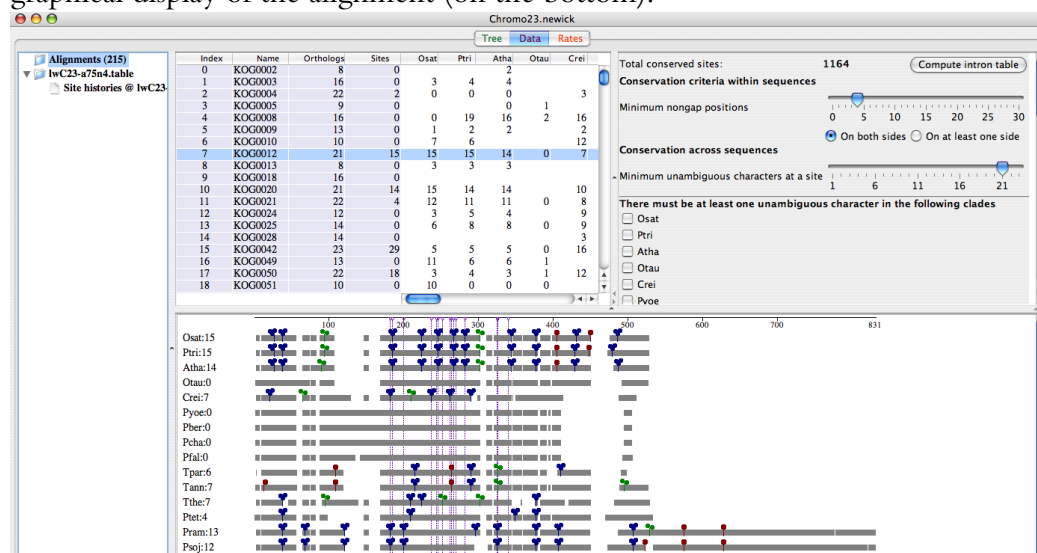
### 2.3.4 Alignment file format: an example

It does not matter where the intron positions and organismal affiliations are specified in the header, as long as they are there.

```
>gnl|Phatr2|15917 /organism=Ftri {i 773,954 i}
-----
-----MENP-----TGAVVPPISLATTTFQQAHPGQATAPEDPNSFGLGYEYSRTGNPT
>gnl|Physo1_1|109222 {i i} unnamed protein /organism=Psoj
-----MSSAASKYVDEHHG
FGTTAIHEGQAPDE-HTGAVAVPITLASTFAQASPGVVAGRGPNSFGKGWEYSRTGNPT
```

### 2.3.5 Alignment panel

The loaded alignments are displayed in an alignment panel, where they can be further processed to identify homologous intron sites. The alignment panel has three parts: a table displaying information about each alignment file (on the upper left), a control panel for specifying conservation criteria (on the upper right), and a graphical display of the alignment (on the bottom).



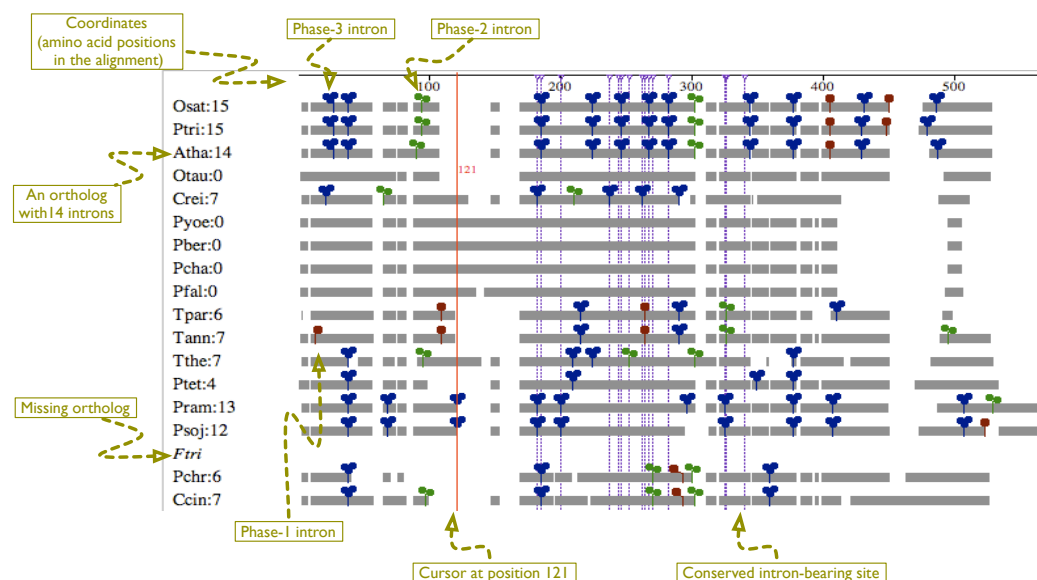
### 2.3.6 Alignment panel: table

Index	Name	Orthologs	Sites	Osat	Ptri	Ati
0	KOG0002	8	0			
1	KOG0003	16	0	3	4	
2	KOG0004	22	2	0	0	
3	KOG0005	9	0			
4	KOG0008	16	0	0	19	1
5	KOG0009	13	0	1	2	
6	KOG0010	10	0	7	6	
7	KOG0012	21	15	15	15	1
8	KOG0013	8	0	3	3	
9	KOG0018	16	0			

The table in the upper left of the alignment panel contains a row for each alignment. The table columns are: Index (0-based indexing of the alignment files), Alignment (file name after stripping the file extension), Orthologs (number of orthologous sequences in the file), Sites (number of conserved intron-bearing sites), and the number of introns for each species.

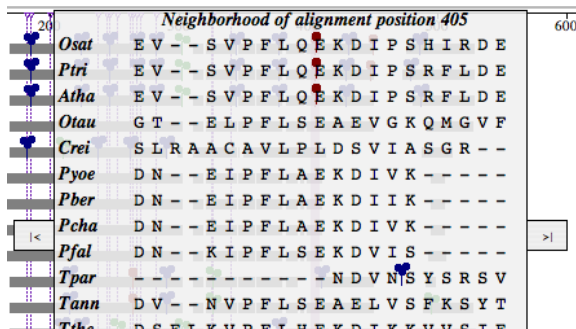
Table rows can be selected one at a time. The alignment for the selected row is illustrated in the bottom panel.

### 2.3.7 Alignment panel: graphical alignment display



The graphical display of an alignment shows the gaps and intron positions with respect to each aligned sequence. Nongap positions are indicated by grey boxes. Intron positions are indicated by small colored tags that have one, two, or three overlapping discs, indicating intron phase. Conserved (i.e., unambiguously aligned) intron-bearing sites are shown by dotted vertical lines. Mouse movements over the alignment display are tracked by a cursor (a red vertical line), indicating the

position within the alignment.



The sequence neighborhood of an intron site can be inspected by double-clicking near the site with the mouse. The appearing overlay panel has two side flaps for the purposes of navigation along the alignment. If you click on a side flap with the mouse, the overlay panel moves to the previous (upstream) or the following (downstream) intron site. The overlay panel disappears for a single mouse click (anywhere outside the side flaps).

### 2.3.8 Alignment panel: conservation criteria

The homology of intron sites can be established in unambiguously aligned regions (Rogozin et al. 2003; Rogozin et al. 2005; Csűrös et al. 2007).

In MALIN, the following procedure is used to establish intron site homology. Each potential splice site (i.e., position between consecutive nucleotides of the coding sequence) is categorized as “ambiguous” or “unambiguous.” Unambiguous positions are counted at each aligned site, and if there are enough of them, then the site is considered “conserved.” Conserved intron-bearing sites are collated to generate an intron table.

Total conserved sites: 410 Compute intron table

**Conservation criteria within sequences**

Minimum nongap positions: 0 5 10 15 20 25 30

☒ On both sides ☐ On at least one side

**Conservation across sequences**

Minimum unambiguous characters at a site: 1 6 11 16 21

There must be at least one unambiguous character in the following clades

- ☐ Osat
- ☐ Ptri
- ☐ Atha
- ☐ Otau
- ☐ Crei
- ☐ Pyoe

A potential splice site is “unambiguous” if it has enough many non-gap amino acid positions in the aligned sequence to the left and right. The requirement may be that  $n$  non-gap positions must be on the left (upstream) and right (downstream), or that  $n$  non-gap positions must be on at least one side<sup>a</sup>:  $n$  is set by a slider, and the requirement option is selected by radio buttons.

<sup>a</sup>for phase-3 introns, there must be at least one non-gap position on each side

The minimum number of unambiguous positions is set by the bottom slider. In addition, clade-specific conservation criteria can be set by using the checkboxes. This latter feature may be useful when one studies intron retainment between or within some specific evolutionary groups.

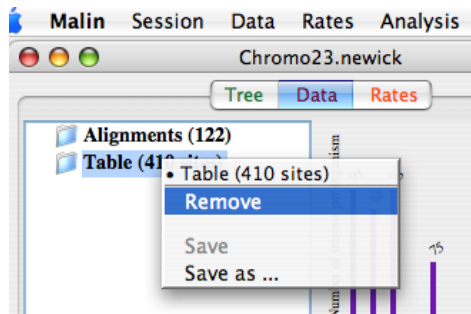
The graphical alignment display and the table are updated as the conservation criteria are changed. At the top of the conservation panel, the number of intron-bearing conserved sites is displayed.

By clicking on the button “Compute intron table”, an intron table is generated from the alignments. A table generated this way contains more information than simple presence-absence: it is also recorded where the sites come from (alignment name and site position with phase), which can be later examined when analyzing site histories.

## 2.4 Intron table

An intron table gives intron presence-absence information in homologous sites across the studied organisms (i.e., terminal taxa of the session phylogeny). An intron table may be generated with MALIN from aligned intron-annotated protein sequences, or a precomputed table can be loaded (Menu: Data → Load table...). All implemented analysis methods use intron tables.





Intron tables can be saved through the popup menu for the corresponding node in the data browser.

### 2.4.1 Intron table: ambiguous characters

The intron table is a set of 0-1 (i.e., absence-presence) sequences, one for each species. MALIN also handles ambiguous characters that can account for cases when orthologs were not found for a given species, or when intron site homology cannot be inferred due to uncertainty in the alignments.

Every further analysis step of the intron table requires the specification of how many ambiguous characters are allowed at a site. This latter is set by the table panel's slider.

### 2.4.2 Intron table file format

The intron table is described in a text file of aligned 0-1 sequences for intron presence-absence in homologous positions. In addition to resolved 0-1 characters, ambiguous intron states can be specified by a question mark ?. Data lines have two fields, separated by TABs: taxon name (must match the name of a terminal taxon in the phylogeny, or else it will be ignored) and an intron presence-absence string. This is the format used by Rogozin et al. (2003), with the addition of encoding ambiguous characters. (Actually, instead of 1, 2 or 3 can also be used to encode intron presence — I implemented this feature with possible future developments in mind involving phase-specific analyses. In the current version, the characters 1,2 and 3 are treated as being identical.)

```
Amel  011001000?
Tcas  010010000?
Ccin  000100010?
Pbla  1?0000100?
Rory  1?0000100?
Hsap  011001000?
```

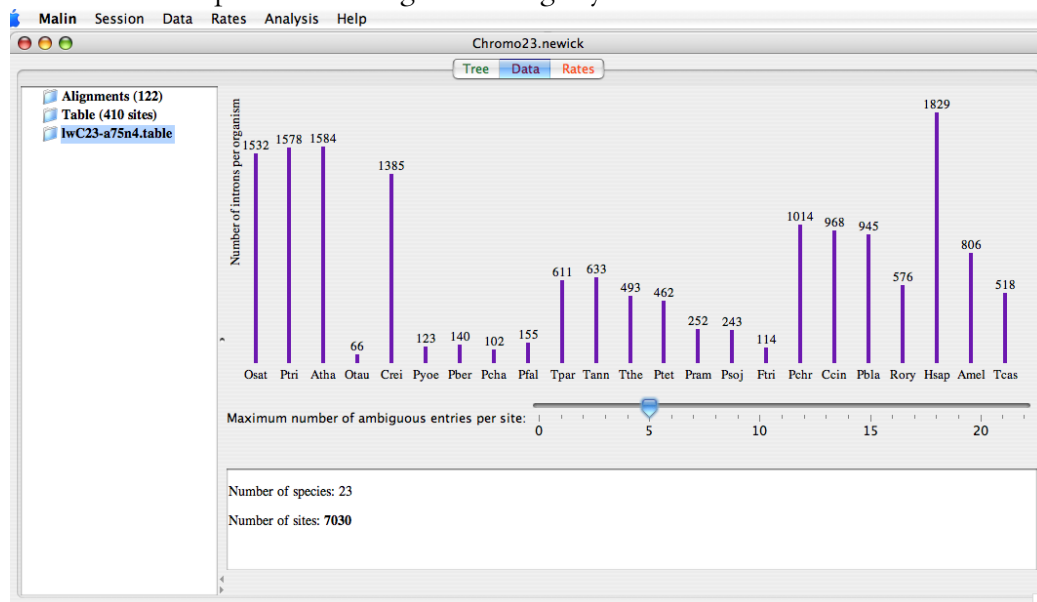
The file may also contain information about mapping the sites to the alignments from which they were computed, in lines that start with #MAP.

```
#MAP    0      KOG0004 1427
#MAP    1      KOG0004 1435
#MAP    2      KOG0004 1455
```

The fields in these lines are: 0-based intron site index, alignment name, and intron site position with respect to the alignment. Intron site positions are 0-based: position 0 is before the first codon, positions 1–2 are phase 1 and 2 intron sites within the first codon, position 3 is between the first and second codons, etc. Other lines in the file that start with # are ignored (supposed to be comments and such).

### 2.4.3 Intron table: table panel

The information panel for an intron table consists of three parts: a bar chart of intron counts per species (on the top), a slider for setting ambiguity thresholds in analysis tasks, and a text area that gives information about the total number of sites satisfying the ambiguity threshold. Note that the bar chart displays the intron counts that are specific for the given ambiguity threshold.



## 2.5 Rates

### 2.5.1 Rate model

MALIN uses a rates-across-sites Markov model for intron evolution, with branch-specific gain and loss rates (Csűrös et al. 2007; Csűrös et al. 2008).

If no rate variation is assumed, then every branch has just a gain and loss rate, with corresponding gain and loss probabilities. Briefly, an intron is lost on an edge with probability

$$p_{\text{loss}} = \frac{\mu}{\lambda + \mu} (1 - e^{-(\lambda + \mu)})$$

where  $\lambda$  and  $\mu$  are the gain and loss rates; a new intron appears in a previously unoccupied site with probability

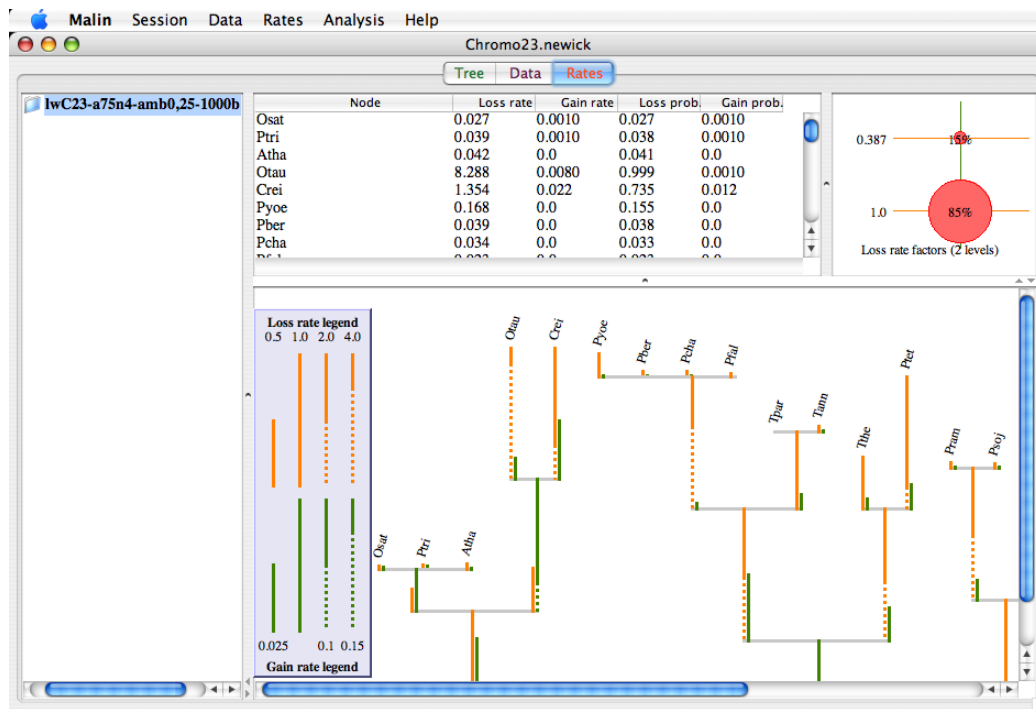
$$p_{\text{gain}} = \frac{\lambda}{\lambda + \mu} (1 - e^{-(\lambda + \mu)}).$$

The constant rate model is completely specified by the branch-specific gain/loss rates, and the *intron density* at the root: this latter is the probability with which intron sites are occupied at the root.

The rate variation model assumes that intron sites belong to discrete loss and gain rate categories. Each site category is defined by a pair of loss and gain *rate factors*  $(\alpha, \beta)$ , so that the loss rates  $\mu\alpha$  and gain rates  $\lambda\beta$  apply on each edge with prototypical loss rate  $\mu$  and gain rate  $\lambda$ .

### 2.5.2 Rates panel

The information panel for a rate model consists of three parts: a table showing numerical values of gain/loss rates and probabilities (on the upper left), a graphical illustration of rate categories (on the upper right), and a graphical illustration of branch-specific gain and loss rates (in the lower half).



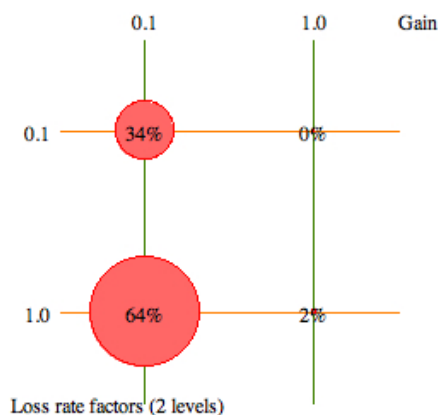
### 2.5.3 Rates panel: table

Node	Loss rate	Gain rate	Loss prob.	Gain prob.
Dm	0.389	0.0040	0.322	0.0030
Ag	0.463	0.0030	0.37	0.0030
Ce	1.087	0.036	0.653	0.022
Hs	0.086	0.032	0.082	0.03
Sp	1.71	0.012	0.816	0.0060
At	0.322	0.065	0.267	0.054
Pf	2.28	0.0030	0.897	0.0010
1 [Diptera]	0.875	0.0060	0.582	0.0040
2 [Ecdysozoa]	0.385	0.0040	0.319	0.0040
3 [Bilateria]	0.064	0.032	0.061	0.03
4 [Opisthokont]	0.022	0.0050	0.021	0.0050
5 [Crown]	1.35	0.0	0.741	5.0E-5
6 [Root]				0.097

The table on the upper left-hand side gives the branch-specific prototypical gain/loss parameters for each branch. Branches are specified by the nodes they lead to. The intron density at the root is shown in the “Gain probability” cell.

If a table row is selected, then the corresponding tree node is highlighted in the graphical display on the bottom.

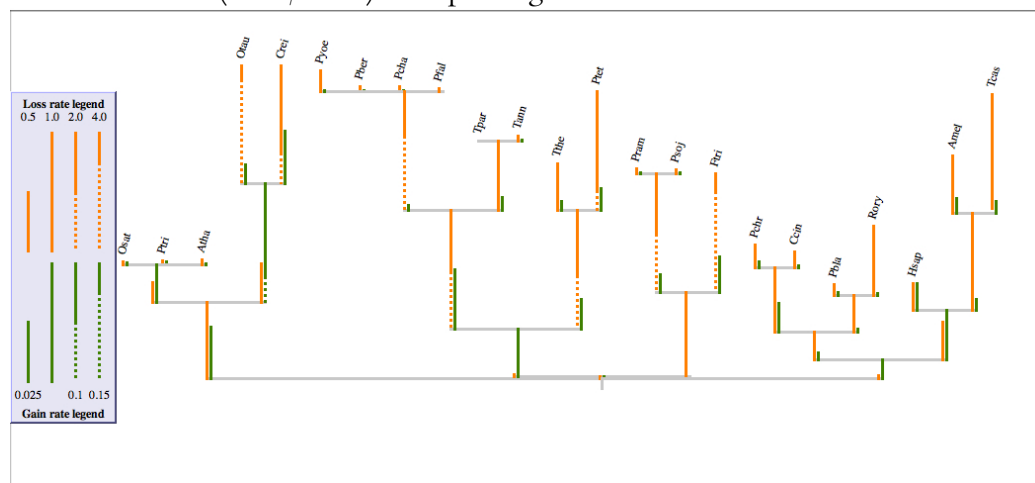
## 2.5.4 Rates panel: graphical display of rate variation



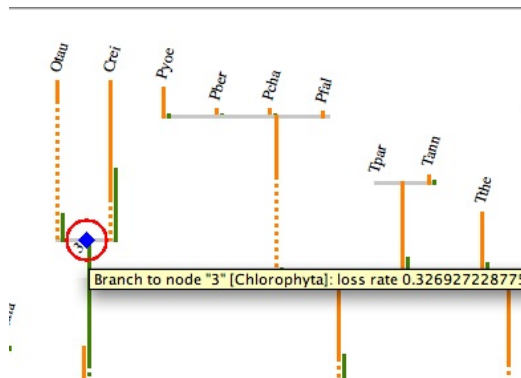
Rate variation is displayed by the grid of loss and gain rate categories. Each site category, defined by the rate factors  $(\alpha_i, \beta_j)$  corresponds to the point at the grid intersection  $(i, j)$ : a disc shows the prior probability for the site category.

## 2.5.5 Rates panel: graphical display of branch rates

The bottom part of the rates panel displays branch-specific loss and gain rates for the neutral class ( $\alpha = \beta = 1$ ) on a phenogram.



Intron loss and gain rates vary by many magnitudes (Roy and Gilbert 2006), and therefore it is not possible to have proportional branch lengths in the display. Instead, an “informative ellipsis” is used for long branches: the ratio between the solid part of the branch and the entire plotted branch length equals the ratio of the displayed branch length and the true branch length. For instance, if the displayed branch length corresponds to a loss rate of  $\mu = 1$ , and the true loss rate is  $\mu = 4$ , then one-quarter of the displayed branch is solid and the rest is dotted.



If a node is selected in the tree (by clicking on it with the mouse), the corresponding row gets selected in the table on the top. When the mouse cursor hovers over a node, an explanatory tooltip pops up.

## 2.5.6 Rate computation

MALIN computes gain and loss rates and other parameters of the probabilistic intron evolution model by the numerical optimization of the likelihood. Prior to proceeding to the actual computation, optimization parameters need to be set in the window that appears after selecting the menu point Rates → Compute rates.

Optimization parameters are grouped under the tabs General parameters, Initial model, Rate variation and Advanced options.

## 2.5.7 Rate optimization: general parameters

On the panel with the tab General parameters, one can set the minima and maxima of branch-specific prototypical gain and loss rates, as well as the convergence criteria for likelihood maximization. The optimization proceeds in rounds — in each

round, the whole set of model parameters are numerically optimized. The optimization stops when the log-likelihood (natural logarithm) changes by less than the given convergence threshold in two consecutive rounds, or after the given number of optimization rounds.

## 2.5.8 Rate optimization: initial model

Numerical optimization starts with a “Rough optimization stage” where parameters of an initial model are set. The initial model may be a default model (with constant rates), or the selected rates model.

Rate optimization

General parameters Initial model Rate variation Advanced options

☐ Default starting model ☒ lwC23-a75n4-amb0,25-1000b.log

Intron presence probability at root: 0.029534956524 ☐ Fixed

Edge to	Loss rate	<input type="checkbox"/> Fixed	Gain rate	<input type="checkbox"/> Fixed
Osat	0.027384082793	<input type="checkbox"/>	0.001073939063	<input type="checkbox"/>
Ptri	0.03912602386	<input type="checkbox"/>	0.001405352277	<input type="checkbox"/>
Atha	0.041990116863	<input type="checkbox"/>	0.000337877213	<input type="checkbox"/>
Otau	8.287988971282	<input type="checkbox"/>	0.008282930073	<input type="checkbox"/>
Crei	1.354285648032	<input type="checkbox"/>	0.022308824208	<input type="checkbox"/>
Pyoe	0.16829358624	<input type="checkbox"/>	0.000299500745	<input type="checkbox"/>
Pber	0.039196191384	<input type="checkbox"/>	0.000198437824	<input type="checkbox"/>
Pcha	0.034022303396	<input type="checkbox"/>	0.000163724851	<input type="checkbox"/>
Pfal	0.02336676027	<input type="checkbox"/>	0.000083831353	<input type="checkbox"/>
Tpar	0.000000001152	<input type="checkbox"/>	0.000210565817	<input type="checkbox"/>
Tann	0.039368059475	<input type="checkbox"/>	0.000618841091	<input type="checkbox"/>
Tthe	0.391797982128	<input type="checkbox"/>	0.003513695516	<input type="checkbox"/>

Perform optimization Cancel

You can also set the starting values for the intron density at the root, as well as branch-specific loss/gain rates. It is also possible to keep some of those parameters fixed throughout the optimization, by checking the corresponding box.

## 2.5.9 Rate optimization: rate variation

Under the Rate variation tab, you can set the number of loss and rate categories for the optimized model, as well as the initial values for the rate factors. The parameters of these categories are optimized in a “Fine optimization stage.”

Rate optimization

General parameters Initial model **Rate variation** Advanced options

Number of loss levels  Number of gain levels

☐ Fixed rate levels

Loss rate factor		Gain rate factor	
0	0.1	0	0.1
1	0.316227766017	1	1
2	1	2	10
3	10		

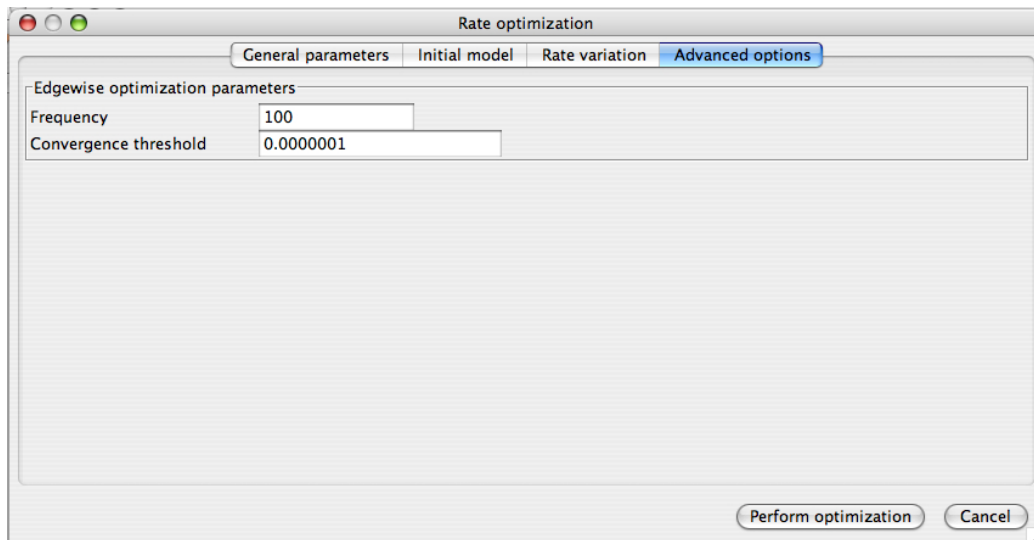
Perform optimization Cancel

It is possible to keep the rate factors fixed during the course of optimization by checking the relevant box on this panel.

### 2.5.10 Rate optimization: advanced options

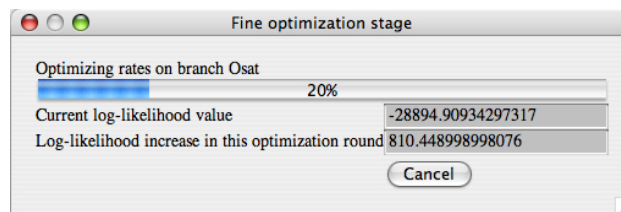
The numerical optimization is carried out by mixing line optimization of individual branch-specific rates (Press et al. 1997) and multidimensional optimization of multiple model parameters at a time (Press et al. 1997). Line optimization is slower, but may improve the likelihood to a larger extent in one optimization round. Under the Advanced options, the frequency of line optimization rounds can be set, as well as the convergence criterion for line optimization.



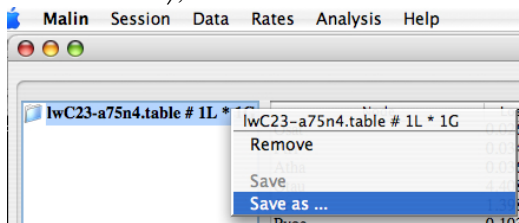


### 2.5.11 Rate optimization: computing

The actual optimization process starts when the Perform optimization button is pressed. The progress of the optimization can be followed in the window that pops up.



At the end of the optimization, either a new model is added to the rates browser (if the default initial model was used, or the rate variation model differs from the initial model), or else the selected rates model is updated in the browser.



The computed rate model can be saved through the popup menu for the corresponding node in the rates browser.

## 2.6 Analysis: shared presence

As the simplest step of data exploration, MALIN can compute the extent to which intron positions at two taxa coincide (Menu: Analysis → Shared presence). Note that shared intron presence is computed for the level of ambiguity set in the intron table panel. The shared presence panel consist of two parts: a table on the top and a graphical display in the bottom.

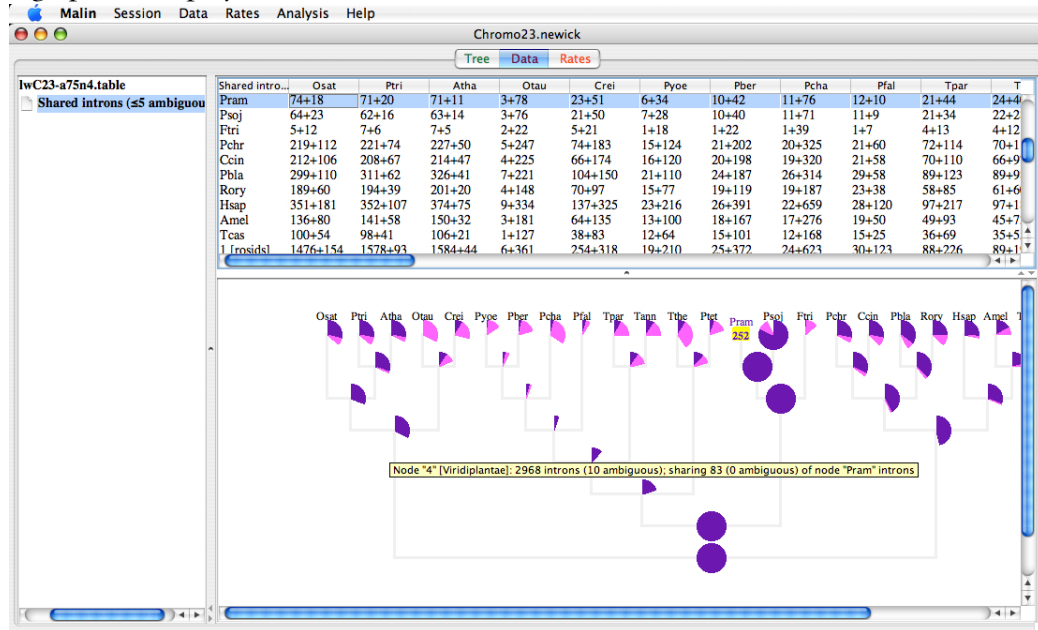


Table rows correspond to the nodes of the session phylogeny. The cell in row  $i$  and column  $j$  tells how many times introns of node  $i$  co-occur with introns of node  $j$ . By definition, an intron “appears” at an inner node, if there is an intron in at least one of the descendant terminal taxa at the given site. The entry  $(i, j)$  is of the form ‘ $a+b$ ’ where  $a$  is the number of sites that contain introns in both  $i$  and  $j$ , and  $b$  is the number of sites that contain an intron in  $i$  but are ambiguous in  $j$ . (For an inner node, a site is considered intron-bearing if at least one descendant of the node contains an intron in that site, and the site is considered ambiguous if at least one descendant is ambiguous, and no descendants have introns.)

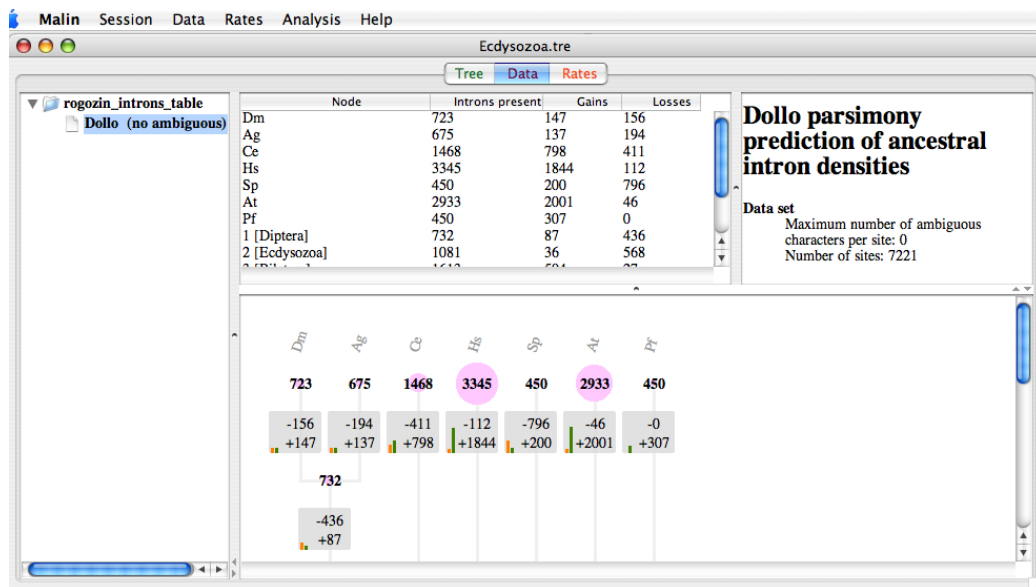
When a row is selected (by a mouse click), the corresponding node is highlighted in the phylogeny on the bottom, and the level of intron sharing is displayed by pie charts at the other nodes. Alternatively, nodes can be selected by clicking on them with the mouse: node selection and table row selection are synchronized. The pie charts show what fraction of intron-bearing sites in the selected node coincide

(dark purple section) or are ambiguous (light purple section) at the other nodes. The numerical values are displayed within a tooltip when the mouse hovers over a node.

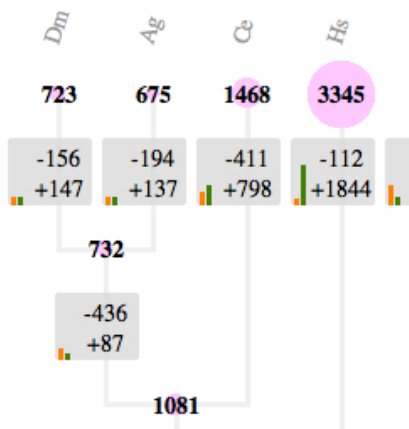
## 2.7 Analysis: Dollo parsimony

MALIN implements the computation of evolutionary history by Dollo parsimony (Menu: Analysis → Evolutionary history by Dollo parsimony). Dollo parsimony (Farris 1977) assumes that gain events are rare. Accordingly, introns at every intron-bearing site are supposed to have a common origin, at the lowest common ancestor of the terminal taxa where the site contains an intron. Intron gain events are rare, and previous studies have used Dollo parsimony (Rogozin et al. 2003; Sullivan et al. 2006) to infer intron ancestry. In my opinion, it is not a good idea, as Dollo parsimony underestimates intron age in the presence of frequent losses, ignores introns that could have been lost in all extant lineages, and does not consider the possibility of parallel intron gains or successive gain-loss-gain events (Csűrös 2005). The original formulation of Dollo parsimony does not allow for ambiguous characters. In the implemented algorithm, the parsimony score is minimized by imposing no penalty to state changes leading to an ambiguous character, i.e., the parent node of a terminal taxon in an ambiguous state may be either in state '0' (absence) or state '1' (presence), without parsimony penalty.

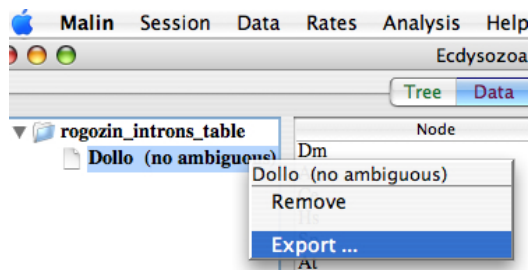
The Dollo parsimony panel consists of three parts: a table (on the upper left), a text area with information about the data (on the upper right), and a graphical tree display (on the bottom).



The table contains the inferred ancestral intron counts, as well as intron gains and losses on tree branches. Each table row corresponds to a node of the session phylogeny. By selecting a row, the corresponding node is highlighted on the bottom. Alternatively, a tree node can be selected on the bottom: node selection and row selection are synchronized.



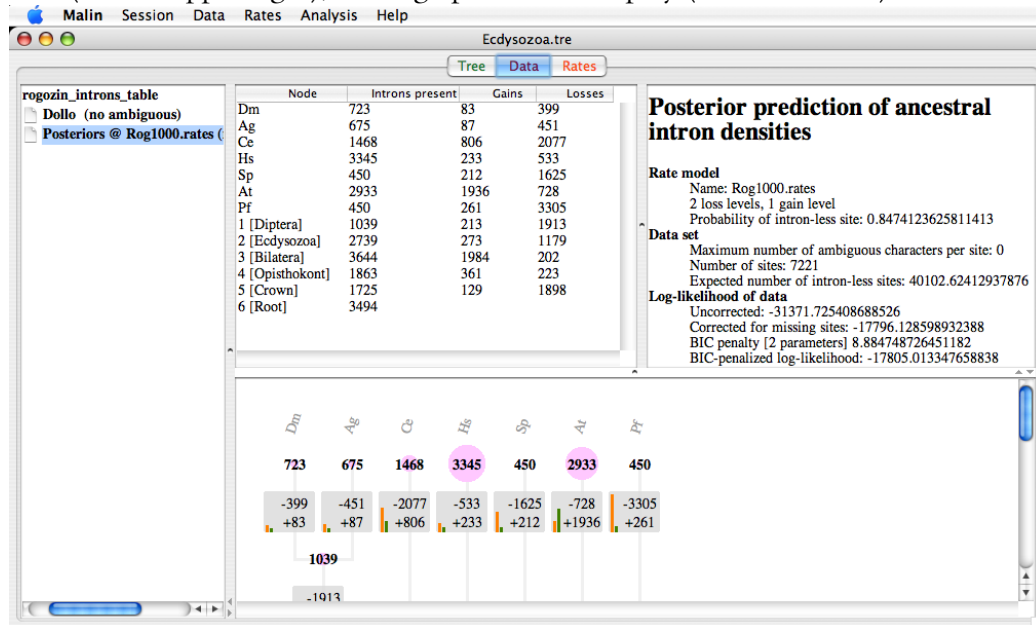
The graphical display on the bottom shows the inferred intron counts at inner nodes and terminal taxa. The purple discs at the nodes have proportional diameters to the inferred densities. On each branch, the inferred number of intron losses (negative numbers) and gains (positive numbers) are shown. The orange and green bars have proportional heights to loss and gain amounts, respectively.



The inferred values can be saved into a tab-delimited text file through the popup menu for the corresponding node in the data browser.

## 2.8 Analysis: posteriors

Probabilistic rate models can be used directly to estimate ancestral intron presence, as well as intron gains and losses (Menu: Analysis → Evolutionary history by posteriors). Namely, these estimates are computed as conditional expectations in the parametric model given the observed data of the intron table (Csűrös et al. 2007). For terminal taxa, the intron counts include the ambiguous sites at which the posterior probability for intron presence is computed in the same probabilistic framework. The information panel for posterior estimates has three parts, just like the Dollo parsimony panel: a table (on the upper left), a text area with information about the data (on the upper right), and a graphical tree display (on the bottom).

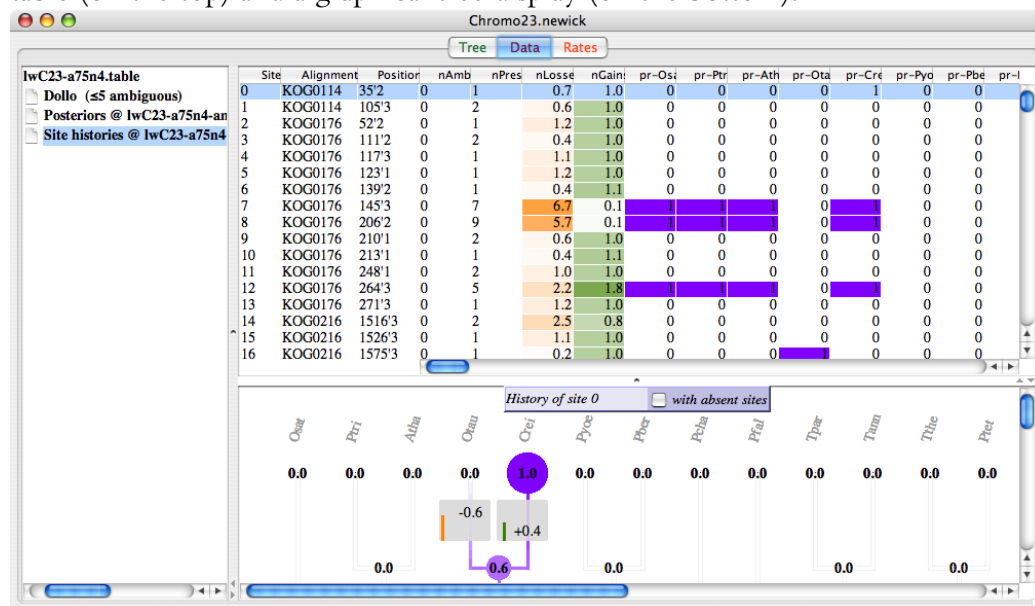


The table and the tree display function in the same way as on the Dollo parsimony panel. The text area, however, is much more informative. First, using the rate

model, MALIN computes the probability with which a site has no intron at any terminal taxon. Since such sites are not included in the intron table, their number is estimated. That number could be interpreted as an indication of the density of potential splice sites (Csűrös 2005; Nguyen et al. 2005), but it is safer to consider it as a numerical characterization of the rate model. Second, the likelihood of the intron table is displayed, either in “uncorrected” form, or by using a correction that accounts for unobserved intron-less sites. The correction method was introduced in the context of restriction site analysis (Felsenstein 1992). The corrected likelihood value can be compared between different probabilistic models to pick the best one. MALIN helps the selection by computing the appropriate model complexity penalty under the Bayesian Information Criterion (Schwarz 1978).

## 2.9 Analysis: site histories

Intron gains and losses at individual or multiple sites can also be analyzed with MALIN (Menu: Analysis → Site histories). The site history panel consists of a table (on the top) and a graphical tree display (on the bottom).



Each table row corresponds to a site in the intron table. If the table was generated from alignments by MALIN, the alignment-specific information about the sites is also available. In such case, the second column in the table gives the alignment name, and the third column gives the alignment position (amino acid position and

intron phase) for the intron site. The graphical display shows the inferred intron gains and losses at sites corresponding to single or multiple selected rows. By double-clicking on a row, all sites belonging to the same alignment are selected. Further table columns are:

**nAmb** number of ambiguous characters at the site

**nPres** number of terminal taxa with an intron at the site

**nLosses** total (expected) number of intron losses at the site (computed as the sum of intron loss probabilities on the branches)

**nGains** total (expected) number of intron gains at the site (computed as the sum of intron gain probabilities on the branches). Notice that **nGains** entries exceeding 1 hint at parallel intron gains.

**pr-*u*** posterior intron presence probability at node *u*

**ls-*u*** posterior intron loss probability on the branch leading to node *u*

**gn-*u*** posterior intron gain probability on the branch leading to node *u*

The table columns can be rearranged (by dragging them with the mouse) to gather those of interest within a visible section of the table.

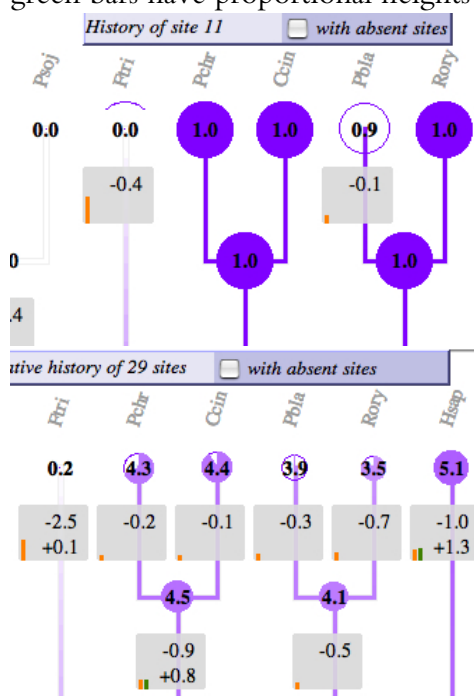
pr-Osi	pr-Ptr	pr-Ath	pr-Ota	pr-Cre	pr-Pyo
1	1	1	0	1	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	1	1	0	0.5	0
1	1	1	0	0.4	1
0	0	0	0	0.0	0
0.0	0.0	0	0.0	0.2	0
0.0	0.0	0	0.0	0.0	0
0.0	0.0	0	0.0	0.0	1

At terminal taxa, ambiguous characters have a fractional presence probability, shown with a decimal point; non-ambiguous characters have integer cell entries (0 for absence, 1 for presence).

Table cells are shaded according to the values they contain (purple for presence, orange for loss and green for gain). Branches of the tree are colored in a similar way with shades of purple.

The graphical display shows intron presence at nodes, and intron losses/gains on the branches in a similar way as the evolutionary history displays (Dollo parsimony and posteriors). When multiple rows are selected, the displayed numbers are the cumulative values for all selected rows. The checkbox allows for the inclusion of inferred intron-less sites, which affects estimates of evolutionary history (see discussion at §2.8). The graphical display on the bottom shows the inferred intron counts

at inner nodes and terminal taxa. The purple discs at the nodes have proportional diameters to the inferred counts. On each branch, the inferred number of intron losses (negative numbers) and gains (positive numbers) are shown. The orange and green bars have proportional heights to loss and gain amounts, respectively.



When a single row is selected, ambiguous characters are indicated by empty discs or arcs (for likely intron absence) at the terminal taxa.

When multiple rows are selected, empty sectors of the discs at the terminal taxa indicate the fraction of ambiguous characters in the selected sites.

## 2.10 Analysis: bootstrap

Uncertainty about inferred rate parameters and evolutionary history can be assessed by bootstrap (Menu: Analysis → Bootstrap for likelihood estimates...). In this procedure, random intron tables (*bootstrap tables*) are generated by selecting sites of the original table independently, with replacement. For each bootstrap table, the likelihood is maximized numerically, and the evolutionary history is calculated by using posterior expectations. The procedure results in a Monte-Carlo approximation of the bootstrap distributions for inferred values (such as loss/gain rates and ancestral densities), which in turn approximate the sampling distributions of the values in question (Efron 1977). In other words, the procedure yields confidence levels for the estimates inferred by MALIN.

The bootstrap procedure is launched after the parameters for likelihood optimization are set in a popup window (see the description of the parameters in the Rate optimization sections). The number of generated bootstrap samples is also set in



this window. Notice that bootstrap samples are generated at the ambiguity level set by the slider in the data display.

**Bootstrap parameters**

Number of bootstrap samples: 100

**Rate optimization parameters**

Maximum number of optimization rounds: 1,000

Convergence threshold on the likelihood: 0.01

Minimum loss rate: 0.000001

Maximum loss rate: 9

Minimum gain rate: 0.000001

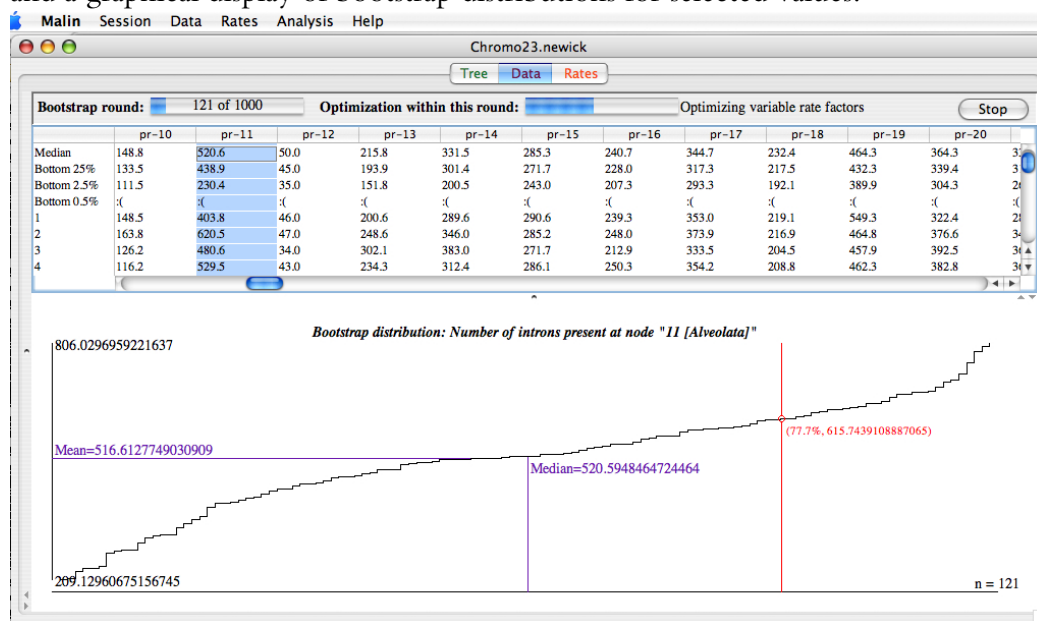
Maximum gain rate: 9

Number of loss levels: 2

Number of gain levels: 1

Starting model: ☐ Default ☒ lwC23-a75n4-amb0,25-1000b.log

Bootstrapping is a time-consuming operation, and it is done in the background. The bootstrap panel consists of a strip of progress monitors (on the top), a table, and a graphical display of bootstrap distributions for selected values.



The progress monitor strip includes a button for interrupting the bootstrap process. The table and the distribution plot are updated continuously as the bootstrapping proceeds.

The top rows of the table show statistics about the bootstrap distributions such as

mean, standard deviation, median, top and bottom percentiles. Additional table rows show the inferred model parameters and estimates in each bootstrap round. Each table column corresponds to a model parameter or estimate. The displayed columns are the following.

**pr- $u$**  Number of introns present at node  $u$  (posterior expectations)

**ls- $u$**  Number of introns lost on the branch leading to  $u$  (posterior expectations)

**gn- $u$**  Number of introns gained on the branch leading to  $u$  (posterior expectations)

**log-likelihood** Natural logarithm of the likelihood

**Root density** Prior probability of intron presence at the root

**Absence prob** Probability that a site has no introns at any terminal taxa

**lsrate- $u$**  Prototypical loss rate on the branch leading to node  $u$

**gnrate- $u$**  Prototypical gain rate on the branch leading to node  $u$

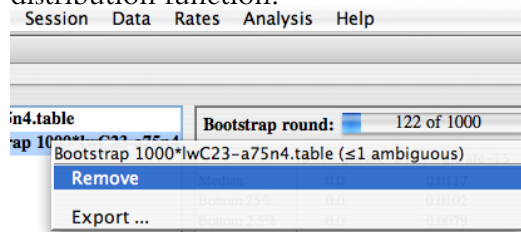
If the rate model implements rate variation, then additional columns are the following.

**Loss # $i$**  Rate factor  $\alpha_i$  for loss category  $i$

**Gain # $i$**  Rate factor  $\beta_i$  for gain category  $i$

**Class  $k$ [ $L_i, G_j$ ]** Prior probability for rate category  $k$  with rate factors  $(\alpha_i, \beta_j)$

Table columns can be rearranged as one sees fit (by dragging them with the mouse). The graphical display on the bottom plots the bootstrap distribution for the selected table column. Along with the cumulative distribution function, the extrema, the median, and the mean are also shown. By moving the mouse over the plot, a cursor can be made to appear (red vertical line), tracking the percentiles in the distribution function.



The inferred values can be saved into a tab-delimited text file through the popup menu for the corresponding node in the data browser.

# Appendix A

## Test data

The distribution includes some test data, packaged in `test.tar.gz`, which expands into files in a `test` directory. The tests include the data set of Rogozin et al. (2003), and another set based on the study of Csűrös et al. (2008) (files starting with `Chromo23`). In particular, the `Rogozin` set consists of the following files.

- `rogozin_introns_table`. The binary intron table provided by the authors.
- `Rogozin-ecdysozoa.tre`. Newick-format tree file showing ecdysozoan phylogeny (common ancestry of insects and worms with the exclusion of vertebrates).
- `Rog-ecdysozoa.rates` and `Rog-ecdysozoa1000.rates`. Rate files with our rate variation, and one loss rate category, respectively.

You can partially repeat the analysis of Roy and Gilbert (2005), by comparing the inference of ancestral intron content using different analysis methods (Dollo parsimony and posterior prediction).

The `Chromo23` set consists of the files

- `Chromo23.newick`. A trifurcating phylogeny for 23 eukaryotic species in the study.
- `Chromo23.table`. Intron table that was used to infer the paper's results.
- `Chromo23.rates`. Rate file with one loss category, which was used to infer the paper's results.

- `Chromo23_alignments`. A directory with 10 out of 394 protein alignments used in the study.

The data set illustrates how site-specific information is dealt with in MALIN, as `Chromo23.table` includes mapping information for intron sites (alignment names, positions and intron phase).

# Appendix B

## Command-line usage

### B.1 Introduction

Some main functions of the MALIN package are accessible from the command line, without launching the graphical user interface. In order to run the command-line program *X*, you need to invoke it through the Java engine:

```
java -cp Malin.jar ca.umontreal.iro.evolution.introns.X...
```

If you run out of memory, you may need to use the `-Xmx` switch as in `java -Xmx512M ...`

The programs write to the standard output, so in order to keep the result, you need to redirect the output into a file, or send it through a pipe to another program. The output typically begins with some information on who, when and how produced the text, written in comment lines that start with `#|`.

### B.2 File types and formats

The command-line programs work with the following main file types.

⟨**tree file**⟩ describes a phylogeny in the Newick format (<http://evolution.genetics.washington.edu/phylip/newicktree.html>) as explained in Section 2.1.

⟨**table file**⟩ The file containing the 0-1 strings of intron presence-absence in homologous positions. The format is discussed in §2.4.2

⟨**rate file**⟩ An intermediate text file listing loss and gain rates along branches. This file is the output of the program `calculateRASRates`.

## B.3 Rate optimization

You can perform rate optimization from the command line, using the program

```
calculateRASRates [switches] ⟨table file⟩ ⟨tree file⟩.
```

The program produces a ⟨rate file⟩ by fitting the parameters (gain & loss rates, length on each branch and root's intron density) to the data to maximize likelihood. The following switches are implemented. (*d* stands for a double-precision floating-point and *i* stands for an integer).

**ambiguous** Specifies the maximum fraction of ambiguous characters allowed in a column. Syntax: `-ambiguous d`. Columns with more than  $d \cdot n$  question marks are ignored in the optimization (*n* is the number of taxa).

**opt-eps** Specifies a stopping rule for the optimization. The optimization stops if the log-likelihood (natural logarithm) changes by less than this value in two consecutive rounds. This switch corresponds to the “Convergence threshold on the likelihood” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-opt-eps d`.

**opt-round** Specifies a stopping rule for the optimization. The optimization stops after that many rounds. This switch corresponds to the “Maximum number of optimization rounds” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-opt-round i`.

**start-with** Specifies that the optimization should start with saved values from a previous execution of `calculateRASRates`. If the starting rate file contains rate variation, then other switches determining the rate variation model structure (`-min-loss-levels`, ..., `max-gain-levels`) are ignored. Syntax: `-start-with ⟨rate file⟩`.

**low-loss-levels** Specifies the number of loss rate factors that are less than 1.0. Syntax: `-low-loss-levels i`.

**high-loss-levels** Specifies the number of loss rate factors that are larger than 1.0. Syntax: `-high-loss-levels i`.

**low-gain-levels** Specifies the number of gain rate factors that are less than 1.0.  
Syntax: `-low-gain-levels i`.

**high-gain-levels** Specifies the number of gain rate factors that are larger than 1.0.  
Syntax: `-high-gain-levels i`.

**min-loss-rate** Specifies the minimum value of the prototypical loss rate on any branch. This switch corresponds to the “Minimum loss rate” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-min-loss-rate d`.

**max-loss-rate** Specifies the maximum value of the prototypical loss rate on any branch. This switch corresponds to the “Maximum loss rate” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-max-loss-rate d`.

**min-gain-rate** Specifies the minimum value of the prototypical gain rate on any branch. This switch corresponds to the “Minimum gain rate” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-min-gain-rate d`.

**max-gain-rate** Specifies the maximum value of the prototypical gain rate on any branch. This switch corresponds to the “Maximum gain rate” field of the rate optimization parameters in the graphical user interface (§2.5.7). Syntax: `-max-gain-rate d`.

**fixrate** Fixes the gain or loss rates on certain branches. Syntax: `-fixrate <constraints>`, where <constraints> comma-separated edgewise constraints. Edgewise constraints have the format <clade>: <rate> or <clade>: <rate>: <rate>, where <clade> is a taxon name in your Newick file (the bottom node for the branch), and <rate> has the syntax *Ld* or *Gd* fixing Loss or Gain rate at *d*. Example: `java -cp Malin.jar ca.umontreal.iro.evolution.introns.calculateRASRates -fixrate 'Diptera:G0:L00.01,Ecdysozoa:G0' rogozin_introns_table Rogozin-ecdysozoa.tre`.

**edgewise-frequency** Frequency of edgewise rate optimization. This switch corresponds to the “Frequency” field of the advanced optimization parameters in the graphical user interface (§2.5.10). Syntax: `-edgewise-frequency i`.

**edgewise-bracket** Bracketing parameter during edgewise rate optimization. This switch corresponds to the “Convergence threshold” field of the advanced

optimization parameters in the graphical user interface (§2.5.10). Syntax:  
`-edgewise-bracket` *d*.

## B.4 Computation of posterior predictions

You can compute the posterior mean predictions for intron counts and intron gains/losses, by using the program

```
computeRASPosteriors [<switches>] <table file> <tree file> <rate file>
```

Currently the only implemented switch is `-ambiguous` in the same syntax as with `calculateRASRates`.

## B.5 Computation of Dollo parsimony

You can compute the Dollo parsimony predictions for intron counts and intron gains/losses, by using the program

```
DolloParsimony [<switches>] <table file> <tree file>
```

Currently the only implemented switch is `-ambiguous` in the same syntax as with `calculateRASRates`.



# Bibliography

- Csűrös, M., I. B. Rogozin, and E. V. Koonin (2008). Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum likelihood approach. *Molecular Biology and Evolution*. Forthcoming.
- Csűrös, M. (2005). Likely scenarios of intron evolution. *Lecture Notes in Computer Science* 3678, 47–60. Proc. RECOMB Satellite Workshop on Comparative Genomics.
- Csűrös, M., J. A. Holey, and I. B. Rogozin (2007). In search of lost introns. *Bioinformatics* 23(13), i87–i96.
- Efron, B. (1977). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7(1), 1–26.
- Farris, J. S. (1977). Phylogenetic analysis under Dollo’s law. *Systematic Zoology* 26(1), 77–88.
- Felsenstein, J. (1992). Phylogenies from restriction sites, a maximum likelihood approach. *Evolution* 46, 159–173.
- Nguyen, H. D., M. Yoshihama, and N. Kenmochi (2005). New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Computational Biology* 1(7), e79.
- Press, W. H., S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery (1997). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge University Press.
- Rogozin, I. B., A. V. Sverdlov, V. N. Babenko, and E. V. Koonin (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Briefings in Bioinformatics* 6(2), 118–134.
- Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin (2003). Remarkable interkingdom conservation of intron positions and mas-

- sive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology* 13, 1512–1517.
- Roy, S. W. and W. Gilbert (2005). Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proceedings of the National Academy of Sciences of the USA* 102(16), 5773–5778.
- Roy, S. W. and W. Gilbert (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* 7, 211–221.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2), 461–464.
- Sullivan, J. C., A. M. Reitzel, and J. R. Finnerty (2006). A high percentage of introns in human genes were present early in animal evolution: Evidence from the basal metazoan *Nematostella vectensis*. *Genome Informatics* 17, 217–229.