# SDI manual

Christian M. Zmasek
Dept. of Genetics
Box 8232
Washington University School of Medicine
4566 Scott Ave.
St. Louis, MO 63110
USA
www.genetics.wustl.edu/eddy/forester/SDI_manual.html
zmasek@genetics.wustl.edu

## 1. Overview

SDI is the name of our algorithm for speciation - duplication inference [1]. This algorithm has a worst-case running time of $O(n^2)$ which is inferior to two previous algorithms that are ~O(n) for a gene tree of n sequences [2,3,5]. However, our algorithm is extremely simple, and its asymptotic worst case behavior is only realized on pathological data sets. We have shown empirically, using 1750 gene trees constructed from the Pfam protein family database, that it appears to be a practical (and often superior) algorithm for analyzing real gene trees [1]. Its Java implementation is part of the FORESTER framework (a framework for automated phylogenomics, available at: http://www.genetics.wustl.edu/eddy/forester/). After installation, the source code for all SDI-related classes is localized in directory "forester/tools/". This also contains a Java implementation of Oliver Eulenstein's algorithm for gene duplication inference [2,3] named "OE.java".

This document describes how to compile and use the various classes which use SDI. In addition, it gives a very brief overview of the classes themselves. For more detailed information, please refer to the API documentation for FORESTER available at: http://www.genetics.wustl.edu/eddy/forester/forester_API_specification/.

FORESTER is based on reading and writing trees in NHX (New Hampshire eXtended) format, which has separate fields for sequence names, species names, duplication - speciation, etc. A more detailed description of NHX is available at: http://www.genetics.wustl.edu/eddy/forester/NHX.html. The programs described in this document can also read trees in regular New Hampshire format, as long as the sequence names are SWISS-PROT names, and therefore contain an abbreviated species name. Obviously, the species tree used needs to follow the same naming convention. A species tree which can be used for the SDI programs and which contains SWISS-PROT style species abbreviations is available at: http://www.genetics.wustl.edu/eddy/forester/tree_of_life_bin_1-3.nhx.

Annotated trees can be displayed with ATV [4], which is part of FORESTER. Newer versions (1.7 or greater) of the ATV application also allow to perform various SDI related operations directly on the displayed gene tree.

For the following commands, it is assumed that the user is in the parent directory for "forester" (but a PATH - and/or alias - can always be set to avoid too much typing).

## 2. Compiling

**Command:**
%javac forester/tools/*.java

Requires at least JDK 1.2

## 3. How to use various classes from the command line ("main" methods)

### 3.1. Duplication inference on a single rooted gene tree - "SDI"

**Command:**
%java forester/tools/SDI [-options] < species tree file name > < gene tree file name > [outfile name]

**Options:**

- -e to use Eulenstein's algorithm [2,3] instead of SDIse
- -n input trees are in New Hampshire format instead of NHX - or the gene tree is in NHX, but species information is only present in the form of SWISS-PROT sequence names (which have the form NAME_SPECIES)

**Species tree file:**
In NHX format, with species names in species name fields - unless -n option is used.

**Gene tree file:**
In NHX format, with species names in species name fields and sequence names in sequence name fields - unless -n option is used.

## 3.2. Duplication inference and (re)rooting on a single unrooted (or rooted) gene tree - "SDIunrooted"

This roots or reroots a gene tree by minimizing a criterion. It allows to root either by minimizing the mapping cost L [5] (also minimizes the sum of gene duplications), the sum of gene duplications, or the tree height (largest distance from root to external node) - the minimizing of which is the same as "midpoint rooting". It also allows to first minimize L or the sum of duplications, and then select out of the resulting trees the one with minimal height (options "-lh" or "-dh"). A certain number (TREES_TO_RETURN) of the resulting gene trees (rooted and with duplication - speciation events assigned) will be displayed using ATV once the analysis is done.

**Command:**
%java forester/tools/SDIunrooted [-options] < species tree file name > < gene tree file name >

**Options:**

- -e to use Eulenstein's algorithm instead of SDIse
- -n input trees are in New Hampshire format instead of NHX - or the gene tree is in NHX, but species information is only present in the form of SWISS-PROT sequence names
- -l to root by minimizing the mapping cost L [5] (and also the sum of duplications)
- -d to root by minimizing the sum of duplications
- -h to root by minimizing tree height (can be used together with -l or -d)

**Species tree file:**
In NHX format, with species names in species name fields - unless -n option is used.

**Gene tree file:**
In NHX format, with species names in species name fields and sequence names in sequence name fields - unless -n option is used.

## 3.3. For duplication inference and (re)rooting on all unrooted (or rooted) gene trees in a directory - "SDIdirectory"

The output of this is a (re)rooted tree with speciation - duplication assigned (in NHX format) for each tree in "gene tree directory" with suffix "suffix for gene trees", as well as a summary list ("outputfile name").

The summary list contains the following. The number in brackets indicates how many external nodes of the gene tree had to be removed since the associated species was not found in the species tree. "en" indicates the number of external nodes in the resulting (analyzed and returned) gene tree. "x" indicates how many differently rooted trees minimized the criterion. "d" are the number of duplications, "L=" the mapping cost, "h=" the height of the resulting, "d=" the minimal difference in tree heights of the two subtrees at the root (this number is 0.0 for a midpoint rooted tree) of the resulting, analyzed and rooted gene tree(s).

The output file ending with "_Sdist" is a file which lists the distribution of trees sizes, "_Ddist" lists the distribution of the sums of duplications (up to a certain maximal size, set with final variables MAX_EXT_NODE_DIST and MAX_DUP_DIST).

**Command:**
%java forester/tools/SDIdirectory [-options] < gene tree directory > < suffix for gene trees > < species tree file name> < output directory > < outputfile name >

**Options:**

- -e to use Eulenstein's algorithm instead of SDIse
- -n input trees are in New Hampshire format instead of NHX - or the gene trees are in NHX, but species information is only present in the form of SWISS-PROT sequence names
- -l to root by minimizing the mapping cost L (and also the sum of duplications)
- -d to root by minimizing the sum of duplications
- -h to root by minimizing tree height (can be used together with -l or -d)
- -w to write assigned gene trees into output directory

**Gene tree directory:**
The directory from which to read gene trees. The gene trees can either be rooted, in which case no rooting with -l, -d, or -h is necessary, or they can be unrooted, in which case rooting is mandatory.

**Suffix for gene trees:**
Suffix of the gene trees to analyze (e.g. "NHX").

**Species tree file:**
In NHX format, with species names in species name fields - unless -n option is used.

**Output directory:**
The directory into which the assigned gene trees will be written.

**Outputfile name:**
File name for summary output files.

# 4. Overview of the classes

## 4.1. SDI (forester/tools/SDI.java)

Abstract class from with classes SDIse and OE inherit. Its main method allows to use/test our SDI (SDIse) algorithm [1] or Eulenstein's algorithm [2,3] for gene duplication inference on a rooted binary gene tree.

## 4.2. SDIse (forester/tools/SDIse.java)

Implements our algorithm for speciation - duplication inference (SDI) [1].

## 4.3. OE (forester/tools/OE.java)

Implements Oliver Eulenstein's algorithm for speciation - duplication inference [2,3].

## 4.4. SDIunrooted (forester/tools/SDIunrooted.java)

Reroots a gene on each of its branches. Performs SDIse or OE on each of the resulting trees. Trees which minimize a certain criterion are returned as the "correctly" rooted ones. The criterions are:

- Sum of duplications
- Mapping cost L [5]
- Tree height, which is the largest distance from root to external node (the minimizing of which is the same as "midpoint rooting")

## 4.5. SDIdirectory (forester/tools/SDIdirectory.java)

Infers duplications - speciations on all (rooted or unrooted) gene trees in a directory by using method "infer" of class SDIunrooted.

## 4.6. BasketDataStructure (forester/datastructures/BasketDataStructure.java)

An implementation of the "Basket" datastructure [3]. This datastructure is used in Eulenstein's algorithm for gene duplication inference ("forester/tools/OE").

## 5. References


[1] Zmasek C.M. and Eddy S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, in press. [MS Word preprint] [software available at http://www.genetics.wustl.edu/eddy/forester/]

[2] Eulenstein, O. (1996) A linear time algorithm for tree mapping. Arbeitspapiere der GMD No. 1046, St Augustine, Germany. [PDF file available at http://taxonomy.zoology.gla.ac.uk/rod/genetree/maths/Linear.pdf]

[3] Eulenstein, O. (1998) Vorhersage von Genduplikationen und deren Entwicklung in der Evolution. GMD Research Series, No 20/1998 GMD - Forschungszentrum Informationstechnik GmbH. - Sankt Augustin. [available at http://www.gmd.de/publications/research/1998/020/]

[4] Zmasek C.M. and Eddy S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17, 383-384. [PubMed] [Bioinformatics] [PDF] [software available at http://www.genetics.wustl.edu/eddy/atv/]

[5] Zhang, L. (1997) On a Mirkin-Muchnik-Smith Conjecture for Comparing Molecular Phylogenies. *Journal of Computational Biology*, 4, 177-187. [PubMed]