# SEADOG-Gen 1.0 Manual

## Description

SEADOG-Gen is a software package for joint phylogenetic reconciliation of domain, gene, and species trees for microbial species. It takes as input a rooted domain tree, one or more rooted gene trees (corresponding to the gene families in which the domains of that domain tree are found), and a rooted species tree for the microbial species considered in the analysis. SEADOG-Gen implements two heuristic algorithms for estimating optimal Generalized Domain-Gene-Species reconciliations, i.e., it outputs a reconciliation of the domain tree with the gene trees and of the gene trees with the species tree while allowing for domain duplication, domain transfer (inter/intra species as well as within/across gene families), domain loss, gene duplication, gene transfer, and gene loss. SEADOG-Gen assumes that both domain transfer and gene transfer occur frequently in the species being analyzed, and the implemented algorithms only work well under this assumption. For Domain-Gene-Species reconciliation in the context of non-microbial species (without frequent horizontal transfer) we recommend using the SEADOG software package instead (https://compbio.engr.uconn.edu/software/seadog/). SEADOG-Gen implements the Generalized Domain-Gene-Species (Gen-DGS) reconciliation model and approximation algorithms introduced in the following manuscript.

*Generalizing the Domain-Gene-Species Reconciliation Framework to Microbial Genes and Domains*
Abhijit Mondal and Mukul S. Bansal.
Under review.

SEADOG-Gen is available open source under GPL version 3. The software is free to use but WITHOUT any guarantee of correctness.

## Input

The required input consists of one domain tree, a set of gene trees, and a species tree. All trees should be in Newick format, without any labels for internal nodes, and should be rooted and binary. Each tree must appear in a separate file. The domain tree and species tree can have arbitrary file names, but gene tree names should be of the form "geneTreeName.tree". All gene trees should be in a single directory. Each leaf label (domain name) in the domain tree must be of the form "domainName_GeneID_GeneTreeName", where the domainName is any label for that specific domain sequence, GeneID is the unique gene ID or name of the specific gene that contains that domain sequence, and GeneTreeName is the name of the gene tree that contains that specific gene. For example, "domainXYZ_FBgn0100324_geneTree4".

Likewise, gene names in the gene trees should be of the form "GeneID_SpeciesName", where GeneID is the unique gene ID or name for that gene, and SpeciesName is the label of the species from which that gene was sampled.

## Command Line Arguments

The program takes the following command line arguments, among which -d, -g, and -s are required and the others are optional.

**List of arguments**:

-d The input domain tree file.

-g Path to the directory containing the gene trees.

-s The input species tree file.

-o Output file name. By default the output file will be the input domain tree file name plus ".output".

-a Algorithm used for computing Gen-DGS reconciliation: 'simpleApprox' for selecting the 'SimpleApprox' algorithm and 'improvedApprox' for selecting the 'ImprovedApprox' algorithm. Default is 'ImprovedApprox' algorithm.

-DD Domain duplication cost. Default value 2.

-DL Domain loss cost. Default value 1.

-DTWW Domain transfer cost when the donor and recipient are in the same gene family and within the same species. Default value 3.

-DTAW Domain transfer cost when the donor and recipient are in different gene family but within the same species. Default value 4.

-DTWA Domain transfer cost when the donor and recipient are in the same gene family but in different species. Default value 4.

-DTAA Domain transfer cost when the donor and recipient are in different gene family and in different species. Default value 5.

-GD Gene duplication cost. Default value 2.

-GL Gene loss cost. Default value 1.

## Command Line Example and Test Data:

```
java -jar Seadog-Gen.jar -d domaintree.tree -g GeneTrees/ -s species.stree -o output.txt
```

The software package includes test data in the "TestData" directory. To execute SEADOG-Gen on this test data, you can use the following command:

```
java -jar Seadog-Gen.jar -d TestData/domain.newick -g TestData/geneTrees/ -s
TestData/species.newick -o testOutput.txt
```

## Output

The reconciliation output begins with a listing of the domain tree, gene tree(s), and species tree, with their internal nodes labeled. Each internal node of the species tree is labeled as 'nx' where x is the pre-order traversal number of the internal node considering only the internal nodes of the species tree. Each internal node of each gene tree is labeled as 'mx_geneTreeName' where x is the pre-order traversal number of the internal node considering only the internal nodes of the gene tree and geneTreeName is the name of the file

containing that gene tree. Each internal node of the domain tree is labeled as 'dx' where x is the pre-order traversal number of the internal node considering only the internal nodes of the domain tree.

The next output block shows the reconciliation of the domain tree with the gene tree(s), showing the event type and mapping for each domain tree node. The reconciliation between the gene tree(s) and species tree appears next, showing the event type and mapping for each node in the gene tree(s).

The output reconciliation ends with three lines showing the final Gen-DGS reconciliation cost, the reconciliation cost between the domain tree and gene tree(s), and the reconciliation cost between the gene tree(s) and the species tree, respectively.

# Contact

If you have any questions, suggestions, or concerns, please contact either Abhijit Mondal (abhijit.mondal@uconn.edu) or Mukul Bansal (mukul.bansal@uconn.edu)