Appendix B - User manual

Contents

1 Input

The needed input for our software is a rooted species tree with exact branch lengths in Newick format and a rooted or unrooted gene tree with exact or inexact branch lengths in Newick format with mapping of gene tree leaves to the species tree leaves.

Inexact branch lengths are required in format, where minimal branch length and maximal branch length are separated with a dash *"-"* such as *"geneName:minimalLength-maximalLength"* shown in an example *"cat:1.3785-2.3375"*.

We recognise two types of leaves mapping. The first type of leaf mapping is directly in the gene tree file, where each gene leaf name consists of the name of the gene and the name of the species to which it is mapping separated with underscore *"_"* such as *"geneName_speciesName"* shown in an example *"cat12_cat"*. The second type of leaf mapping is stored in an *".smap"* file, where each gene name has assigned a species. It is a pattern-matching mapping and we consider only matching at prefix. For example, if the gene name starts always with *"CAT"*, the mapping is *"CAT* cat"*, where the *"*"* is optional postfix after the matching pattern and the gene leaf in a gene tree is *"CAT12:0.8324"*.

2 Output

The output of our software is either reconciled gene trees file or a relations file. For the reconciled gene trees file, we print all reconciliation solution with the same most parsimonious reconciliation score. Each gene tree is printed in a Newick format on a new line.

For the relations file, we recognise 4 types of relations for each node in gene tree:

- *"gene"* - for a gene node in a leaf,
- *"spec"* - if the evolutionary event occurred on the gene node is speciation,
- *"dup"* - if the evolutionary event occurred on the gene node is duplication,
- *"loss"* - if the evolutionary event occurred on the gene node is gene loss.

Each relation type is written in a new line with information about the node separated with spaces, such as:

- *"gene geneName"*, where *geneName* is for the name of the gene
- *"spec leftChild rightChild"*, where *leftChild* is the name of a left child of the speciation node and *rightChild* is the name of a right child of the speciation node,
- *"dup leftChild rightChild"*, where *leftChild* is the name of a left child of the duplication node and *rightChild* is the name of a right child of the duplication node,
- *"loss node"*, where *node* is the name of a node above which the loss occurred.

Each name of an internal node of a gene tree consists of genes presented in a subtree of that node separated by commas *","*. If the found solutions have the same relations, we print only the first and the number of reconciled trees with the same relations.

3 Settings

The mandatory setting for running the isometric software are:

`-S <species tree>`

The *<species tree>* signifies the path to the rooted species tree file in Newick tree format.

`-G <gene tree>`

The *<gene tree>* signifies the path to the rooted or unrooted gene tree in Newick tree format.

If the given gene tree does not have leaf mapping directly int the gene tree file (as we discussed in Chapter 1), we require a mapping file:

`-M <species map>`

The *<species map>* signifies the path to the mapping file of genes to species in format mentioned above in Chapter 1.

Besides the mandatory setting, we offer optional settings:

`-help`

The setting shows help information with all possible input arguments.

`-t <tolerance>`

The *<tolerance>* signifies the required tolerance from interval $[0, 1]$. By default, the tolerance is set to 0.5.

`-s <step>`

The *<step>* signifies the required step from $[0, \infty]$. By default, the step is set to 0.01.

`-r`

The setting refers that the given gene tree is rooted. By default, the given gene tree is considered to be unrooted.

`-reroot`

The setting refers that the given rooted gene tree is wished to be rerooted, which allows to find a reconciliation with better score.

`-l`

The setting refers that the user wishes to count the gene losses above the root of the given gene tree. The case when the gene losses are above the root can occur when some genes of species from the species tree are not presented in the gene tree.

`-p <print type>`

The *<print type>* signifies the required print type from two options *"sol"* and *"rel"* as discussed in Chapter 2. The *"sol"* option prints the reconciled gene trees with the same most parsimonious reconciliation score. The *"rel"* option prints the relations between nodes. By default, the print type is set to be *"sol"*.

`-epsilon <epsilon>`

The *<epsilon>* signifies the required epsilon to recognize the rounding error. By default, the ϵ is set to 1×10^{-6}.

4 Running the reconciliation

Our isometric reconciliation software has a command line interface, so it is executed from the command line with:

```
java -jar IsometricRecon.jar [settings]
```

The *[settings]* are setting discussed in Chapter 3. If the mandatory settings are not inputted or the given trees or leaf mapping is incorrect, the software writes an error message into the command line.