ARTra (Version 1.0)

https://compbio.engr.uconn.edu/software/ARTra/

Description

ARTra (short for "Additive and Replacing Transfer Inference") is a program for inferring and distinguishing between additive and replacing horizontal gene transfer events. ARTra uses Duplication-Transfer-Loss (DTL) reconciliation to infer transfer events and then uses a trained machine learning classifier to classify the inferred transfers as additive or replacing. It also implements three simple non-machine-learning-based classification heuristics, including the "*gene-frequency*" heuristic described in the paper available from https://doi.org/10.1145/3307339.3342168 and implemented in the RANGER-DTL-RT tool. The other two heuristics, "*lost-gene*" heuristic and "*mapping-count*" heuristic, are described in the paper cited below. The machine learning classifier uses the classifications generated by these heuristics, along with some additional features, to generate an improved ensemble classification. Further technical details appear in the paper cited below.

ARTra can be cited as follows:

> *A Supervised Machine Learning Approach for Distinguishing Between Additive and Replacing Horizontal Gene Transfers*
> Abhijit Mondal, Misagh Kordi, Mukul S. Bansal
> ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB) 2020: to appear.

ARTra is freely available from https://compbio.engr.uconn.edu/software/ARTra/

Dependencies

To execute ARTra, users must have the following installed on their systems: Java 8 and Python 3, along with Python's scikit-learn, joblib, matplotlib and pandas libraries.

Usage

ARTra takes as input a single file containing a rooted species tree on the first line and a rooted gene tree on the second line. Both trees must be in Newick format, and leaf labels in the gene tree should be of the form "<SpeciesName>_<GeneName>"; this enables mapping of the leaves of the gene tree to the leaves of the species tree. A sample input file is available in the software directory as input.newick.

Users can also specify an output file name and the specific classification heuristic to be used. If an output file name is not specified then the output is written to the terminal. If a heuristic is not specified then the machine learning classifier is used by default.

ARTra can be executed as follows:

```
java -jar ARTra.jar -i inputFile -o outputFile -h heuristic [-options]
```

Available command line options are listed and described below.

List of command line options

-i Input file name. File should contain species tree on the first line and gene tree on the second line. This is a required parameter.

-o, Output file name. If an output file is not specified then output is written to the terminal.

-h Classification method to be used. Options are “lostGene2”, “lostGene3”, “geneFrequency”, “mappingCount”, and “ml”, with “ml” as the default. Their meanings are as follows:

lostGene1: refers to *lost-gene(h=1)* heuristic.
lostGene2: refers to *lost-gene(h=2)* heuristic.
lostGene3: refers to *lost-gene(h=3)* heuristic.
geneFrequency: refers to *gene-frequency* heuristic.
mappingCount: refers to *mapping-count* heuristic.
ml: refers to the machine learning heuristic and is the default option.

-D Duplication cost (whole number only, default value 2).

-T Transfer cost (whole number only, default value 3).

-L Loss cost (whole number only, default value 1).

Interpretation of the output

The output from an execution of ARTra is written to the specified output file. This output format is nearly identical to that of RANGER-DTL (https://compbio.engr.uconn.edu/software/RANGER-DTL/) and we refer the reader to the RANGER-DTL manual for a detailed description. Briefly, the output consists of copies of the species tree and gene tree with labeled internal nodes, followed by a reconciliation block showing an optimal DTL reconciliation of the gene tree with the species tree. Each transfer node in this reconciliation is labeled as either “additive transfer” or “replacing transfer”. The format of this reconciliation block is identical to the reconciliation output format of RANGER-DTL, except that transfers are additionally labeled as additive or replacing. This reconciliation block is followed by some additional basic information about the computed reconciliation.

Example input and output file

The software directory includes a sample input file (input.newick; containing a species tree and gene tree) and the corresponding output file (output.txt; containing the reconciliation output with classified transfers). The software can be executed on this input file using the following command:

```
java -jar ARTra.jar -i input.newick -o classifierOutput.txt -h ml
```

Retraining the machine learning classifier

The software includes a pretrained classifier (file named “saved_classifier.joblib”). However, it is easy to retrain the classifier, if needed, using additional or custom training data. The actual training file used to train the default classifier is provided as “all_training.txt” in the software package. Any custom training

file must follow the same CSV data format used in this included training file. The columns in the training file are, in order:

id
height
mappingCountHeuristic
lostGene2Heuristic
lostGene1Heuristic
lostGene3Heuristic
geneFrequencyHeuristic
trueLabel

The definitions of these columns are as follows:

- id: The identifier or name of the transfer node in the gene tree
- height: Height of the transfer node (i.e., number of edges between transfer node and its furthest leaf descendant)
- mappingCountHeuristic: The classification result for that transfer node from the *mapping-count* heuristic. '0' means replacing transfer and '1' means additive transfer.
- lostGene2Heuristic: The classification result for that transfer node from the *lost-gene(h=2)* heuristic. '0' means replacing transfer and '1' means additive transfer.
- lostGene1Heuristic: The classification result for that transfer node from the *lost-gene(h=1)* heuristic. '0' means replacing transfer and '1' means additive transfer.
- lostGene3Heuristic: The classification result for that transfer node from the *lost-gene(h=3)* heuristic. '0' means replacing transfer and '1' means additive transfer.
- geneFrequencyHeuristic: The classification result for that transfer node from the *gene-frequency* heuristic. '0' means replacing transfer and '1' means additive transfer.
- trueLabel: The true label for that transfer event. '0' means replacing transfer and '1' means additive transfer.

Loading and dumping classifiers for reuse: The trained classifier is automatically loaded from the file "saved_classifier.joblib". If that file is absent, the program will attempt to train a new classifier from the training file named "all_training.txt" present in the same directory.

To retrain the classifier using custom training data, the new training file should be named "all_training.txt" and placed in the same directory as "DTRL_Classifier.py" (replacing the original "all_training.txt" file). The existing file named "saved_classifier.joblib" should also be removed so that the program is forced to create a new classifier which would then be trained using the custom training data.

Contact Information

In case of any questions, please feel free to contact Abhijit Mondal (abhijit.mondal@uconn.edu) or Mukul Bansal (mukul.bansal@uconn.edu).