****

Introduction

****

COSPEDTree is a python based tool for computing supertree from input candidate source trees. Supertree computation is formed using a greedy approach of partitioning the input set of taxa based on equivalence relation. The relationship among a set of taxa is determined by ancestor / descendent / sibling / no relation characteristics. Based on such equivalence partitioning, a Directed Acyclic Graph (DAG) is formed initially, from which the output tree is generated.

Input source trees can be either in NEWICK format or in NEXUS format.
However, all the source trees should have identical input formats.
Output tree is generated in the NEWICK format.

****

Dependencies

****

COSPEDTree is developed in Linux Systems (Ubuntu 12.04), using Python 2.7.

If a system alreay has python installed and the package has version lower than 2.7, then there can be problems, as tested. So, in such case, corresponding package upgrade is requested.

The executable is tested with different versions of Ubuntu. For systems having Ubuntu with lower versions, please notify in case of any errors.

The executable does not require any libraries to be installed prior execution.

****

Execution

****

COSPEDTree is provided as a stand along executable.
The executable contains binaries of the source codes and the used libraries (static libraries).
COSPEDTree is to be executed with the following command line options, from a terminal:
(assuming the present working directory contains the executable)

```
./COSPEDTree [options]
```

****

NOTE:

All the options except the first three, signify toggle / complement of their corresponding DEFAULT values.
First option (help) displays these command line parameters.

It Is Preferable For A Beginner, To Not Use Any Option Other Than The Second And Third Options.
Second option is for specifying the input filename (mandatory)
Third option is for specifying the corresponding file format (mandatory for nexus file format data).
****

Details of the options are mentioned below:

```
-h, --help           show this help message and exit

-I INP_FILENAME, --INPFILE=INP_FILENAME
                name of the input file containing candidate source trees

-p INP_FILE_FORMAT, --inpform=INP_FILE_FORMAT
                1 (default) - input file format is NEWICK
                2 - input file format is NEXUS

-q NO_OF_QUEUES, --queues=NO_OF_QUEUES
                1 - only a single max priority queue is used for
                storing the score metrics
                2 (default) - two separate queues are used to store the conflicting
                and non conflicting taxa pairs and corresponding score metrics

-a, --all        if True then for conflicting taxa pairs (at least two
                relations between them are supported by the source
                trees), all possible four relations between them are considered and
                included in the score metric priority queue

                Otherwise if it is False, information for only the
                relations which are supported by the source trees
                are included in the score metric priority queue

                - Default FALSE

-n, --preservenoedge  if true, then it prioritizes NO edge type -
                     once an edge is marked as NO RELATION edge, it will not be changed
                - Default TRUE

-e, --equivpart      if true, then it clusters a group of taxa on the basis
                of equivalence partition
                - Default TRUE

-r, --rooted         if true, then trees are read and processed as rooted trees
                - Default FALSE

-u, --underscore      if true, then this option preserves the underscores of
                the names of taxa
                - Default TRUE

-c, --costupdate      if true, then this option updates the edge costs
                during each iteration of edge connectivity
                - Default TRUE

-s, --singleedgepriority
                if true, then this option connects two taxa in the final supertree
                if those two taxa have only one single relation in the source trees
                that is, non-conflicting taxa pairs are separated first
                - Default TRUE
```

```
 -t, --tiecase       if true, then during selection among multiple edges
                   having equal cost, this option prioritizes certain edge types                   -
                   Default TRUE

 -i, --initcost      if true, then this option uses one single edge cost
                   assignment at the beginning to all the edges subsequently no update
                   of edge costs is performed, and
                   edge selection is carried out with the initial cost settings
                   - Default FALSE

 -f, --fractwt       if true, this option uses fractional values of edge
                   weights for cost updation and edge selection
                   - Default FALSE
```

****
EXAMPLE OF COMMANDS
****

****
CASE A – when user makes dynamic scoring OFF
only static initialized score values are used for supertree construction
corresponding results are provdided in the manuscript
COuplet Supertree by Equivalence Partitioning of taxa set and DAG formation Sourya Bhattacharyya and Jayanta Mukhopadhyay, Proceedings of 5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB), Newport, California, September 2014, pp. 259-268.
****

```
./COSPEDTree -I 'source_tree_input_filename' -c -p 'inp_file_format' -q 'no_of_queues'
```

command descriptions:
1) Using -I command we specify the input filename (denoted by 'source_tree_input_filename').
Replace the filename with the full or relative path of custom input file containing the source trees.

2) -c option is to disable the updation of the score metric values at each iteration.
Score metric values are initialized with the product of priority metric and the frequency
associated with corresponding relation between concerned taxa pair.
Subsequently iterations are carried out with the score metric values.

3) -p option is for specifying the input tree format
(as denoted by 'inp_file_format')
if input file contains the trees in NEWICK format, then specify the option as (-p 1) (1 stands for newick)
if input file contains the trees in NEXUS format, then specify the option as (-p 2) (2 stands for nexus)

4) -q option is for specifying the number of priority queues (by 'no_of_queues')
employed for storing and manipulation of the score metric.
if number of queues is 1 (Q_one in the manuscript), then specify the option as (-q 1)
if number of queues is 2 (Q_two in the manuscript), then specify the option as (-q 2)

****

FINAL OUTPUT SUPERTREE

****

In the same directory containing the source trees (as specified by 'source_tree_input_filename'), upon execution of above command,
one folder will be created as per the following naming convention:

inpfilefmt_'$I'_costupdate_'$c'_no_of_queue_'$q'_include_all_multi_reln_'$a'

Here '$I' corresponds to the value of input file format. It is 1 or 2
'$c' is False since dynamic cost update is disabled.
'$q' is the number of queues. It will be 1 or 2
'$a' is False.

Within the folder, two text files will be created:
1) 'complete_output_description.txt' containing the details of execution,
output supertree and performance metric values (with respect to sumFP, sumFN, and sumRF)
2) Text file "output_supertree_newick.tre" which contains the derived supertree
(in both newick string representation, as well as a tree plot)

The tree can be used subsequently for performance metric computation

****

CASE B – when user makes dynamic scoring ON
corresponding results are provdided in the manuscript
COSPEDTree: COuplet Supertree by Equivalence Partitioning of taxa set and DAG formation Sourya Bhattacharyya and Jayanta Mukherjee, accepted for publication in IEEE/ACM Transaction on Computational Biology and Bioinformatics.

****

```
./COSPEDTree -I 'source_tree_input_filename' -p 'inp_file_format' -q 'no_of_queues'
```

Here there is no -c option, thus enabling the cost update option.
Note: cost update operation is time consuming. So users are advised to rather use the earlier command line option.

Details of output files are same as above.

****

For any queries, please contact

****

Sourya Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
<sourya.bhatta@gmail.com>