PhyloRelief Documentation

Release 1.1

Davide Albanese, Claudio Donati

November 12, 2014

CONTENTS

PhyloRelief implements an algorithm that introduces the Relief strategy of feature weighting in a phylogenetic context to identify those OTUs or groups of OTUs that are correlated for the differentiation between classes of samples (i.e. healthy vs. disease, lean vs. obese etc.) in a metagenomic dataset. By integrating the phylogenetic relationships amongst taxa into the framework of statistical learning, the method is able to unambiguously group the taxa into lineages without relying on a precompiled taxonomy, and accomplishes a ranking of the lineages according to their contribution to the sample differentiation.

CHAPTER

ONE

INSTALLATION

1.1 Requirements

- Linux or OSX systems
- Python >= 2.7
- NumPy and SciPy (http://scipy.org)
- Pandas (http://pandas.pydata.org
- DendroPy (http://pythonhosted.org/DendroPy/)
- Statsmodels (http://statsmodels.sourceforge.net/)

On Linux Required Python modules can easily installed by running the package manager or using `easy_install`:

```
$ sudo easy_install numpy scipy pandas dendropy statsmodels
```

On OSX

1. Install the gfortran compiler from http://cran.r-project.org/bin/macosx/tools/ (required by SciPy)
2. Install the Python modules by running `easy_install`:

   ```
   $ sudo easy_install numpy scipy pandas dendropy
   ```

1.2 Installing PhyloRelief

1. Untar `phylorelief-X.Y.Z.tar.gz`, creating `phylorelief-X.Y.Z` folder (where `X.Y.Z` is the current version of phylorelief)
2. Go into `phylorelief-X.Y.Z` folder and from a terminal run:

   ```
   $ sudo python setup.py install
   ```

3. If you don't have root access, installing phylorelief in a directory by specifying the `--prefix` argument. Then you need to set tht `PYTHONPATH` environment variable:

   ```
   $ python setup.py install --prefix=/path/to/modules
   $ export PYTHONPATH=$PYTHONPATH:/path/to/modules/lib/python{version}/site-packages
   ```

4. Test the installation:

```
$ phylorelief
usage: phylorelief otu_table tree sample_data target [options]
phylorelief: error: too few arguments
...
```

CHAPTER

TWO

USAGE

PhyloRelief is distributed as a command line application. You can run `phylorelief --help` in order to show a summary of options:

```
$ phylorelief --help
usage: phylorelief otu_table tree sample_data target [options]

PhyloRelief v1.1. Phylogenetic-based Relief for clade weighting
and OTU ranking.

positional arguments:
  otu_table             an OTU table file (containing the number of sequences
                        observed in each OTU for each sample) in tab-delimited
                        format
  tree                  a (rooted) tree in 'newick' or 'nexus' format (see
                        --tree-format option). Note that all the leaf nodes
                        should have univocal names
  sample_data           a tab-delimited file containing sample information and
                        metadata
  target                sample data column to be used as class label

optional arguments:
  -h, --help            show this help message and exit
  -v, --version         show program's version number and exit
  -f {newick,nexus}, --tree-format {newick,nexus}
                        tree format (default newick)
  -u {unweighted,weighted,generalized}, --uf-variant {unweighted,weighted,generalized}
                        unifrac variant (default weighted)
  -k N_NEAREST_NEIGHBORS, --n-nearest-neighbors N_NEAREST_NEIGHBORS
                        number of nearest neighbors (default 2)
  -a ALPHA, --alpha ALPHA
                        alpha for generalized unifrac (default 0.5)
  -i N_ITERATIONS, --n-iterations N_ITERATIONS
                        number of iterations (default number of samples)
  -o OUT, --out OUT     clade ranking output file (default
                        out_phylorelief.txt)
  -t TREE_OUT, --tree_out TREE_OUT
                        annotated tree (BEAST/FigTree style) in nexus format
                        (default tree_phylorelief.tre)

Example:

    $ phylorelief otu_table.txt tree.tre sample_data.txt Status -k 1

Authors:
```

```
    Davide Albanese <davide.albanese@fmach.it>
    Claudio Donati <claudio.donati@fmach.it>

Fondazione Edmund Mach, 2013
```

`phylorelief` application takes four required inputs, three files and a target name:

`tree` A (rooted) tree in `newick` format.

Note: Note that all the leaf nodes should have univocal names.

`otu_table` An OTU table (containing the number of sequences observed in each OTU for each sample) in **tab-delimited** format, OTU (rows) x sample (columns) orientation:

OTU	SampleName1	SampleName2	SampleName3	...
OTU1	22	3	6	...
OTU2	10	45	340	...
...	...	...	...	...

OTU names must correspond to the leaf names in the tree.

Note: A rarefied OTU table should be used in order to remove sample heterogeneity.

`sample_table` A **tab-delimited** file containing sample information and metadata (e.g. health status):

Sample	Status	Treated	...
SampleName1	Case	Yes	...
SampleName2	Control	No	...
SampleName3	Case	No	...
...	...	...	...

Sample names must correspond to the sample names in the OTU table.

`target` A `sample_table` column to be used as class label (e.g. 'Status' or 'Treated' in the example above)

The application outputs two files, a clade ranking file and a annotated tree in nexus format:

Clade ranking file

A clade ranking file is a tab-delimited file with five columns. The clades are ranked in decreasing order according the weight assigned by the algorithm. The first column contains the PhyloRelief weights, the 2nd, 3rd and 4th columns contain the statistics, the p-values and the FDR corrected p-values of the Kruskal-Wallis test and the last column contains the OTU names (comma separated) forming the corresponding clade:

phylorelief_weight	kw_stats	kw_p	kw_padj	OTUs
0.8633	0.8597	0.0225	0.0589	OTU34,OTU2,OTU8
0.6402	6.2530	0.0012	0.0242	OTU12,OTU11
...	...	...	...	...

Annotated tree file An annotated tree (BEAST/FigTree style) in nexus format. In this file **each node** in the tree is annotated with two metadata: `phylorelief_weight` and `phylorelief_rank`. `phylorelief_weight` contains the weights assigned by phylorelief to the clade starting from the node and `phylorelief_rank` contains the clade ranking position.

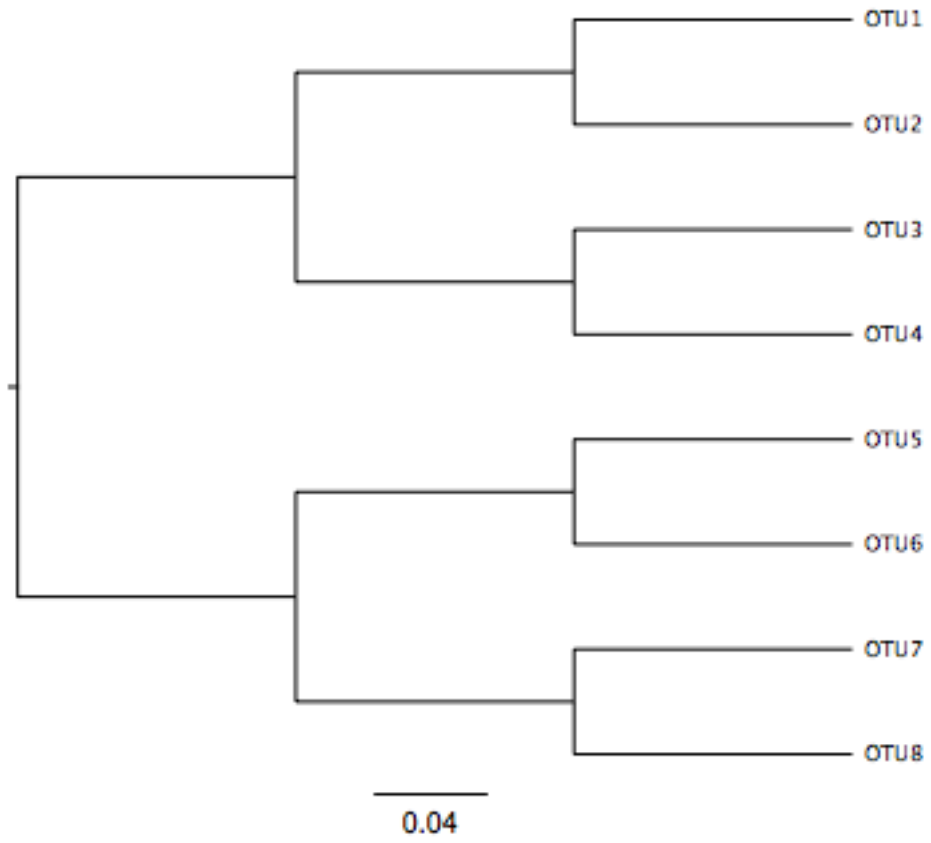
CHAPTER

THREE

EXAMPLE

This very simple example is located in the `example/` directory. The directory contains three files. A rooted tree in newick format:

```
(((OTU1:0.1, OTU2:0.1):0.1,(OTU3:0.1, OTU4:0.1):0.1):0.1,((OTU5:0.1, OTU6:0.1):0.1,(OTU7:0.1,OTU8:0.1
```



An OTU table in tab-delimited format:

OTU	SAMPLE1		SAMPLE2		SAMPLE3	SAMPLE4
OTU1	1	1	2	1		
OTU2	1	2	1	1		
OTU3	1	1	2	2		
OTU4	1	1	2	2		
OTU5	0	0	3	0		
OTU6	0	0	1	4		

```
OTU7        3        3        0        0
OTU8        3        3        0        0
```

A sample table with sample information:

```
Sample        Status
SAMPLE4        Control
SAMPLE1        Case
SAMPLE3        Control
SAMPLE2        Case
```

To to run example, open a terminal, go into the example (examples/) folder and run:

```
$ phylorelief otu_table.txt tree.tre sample_data.txt Status -k 1
```

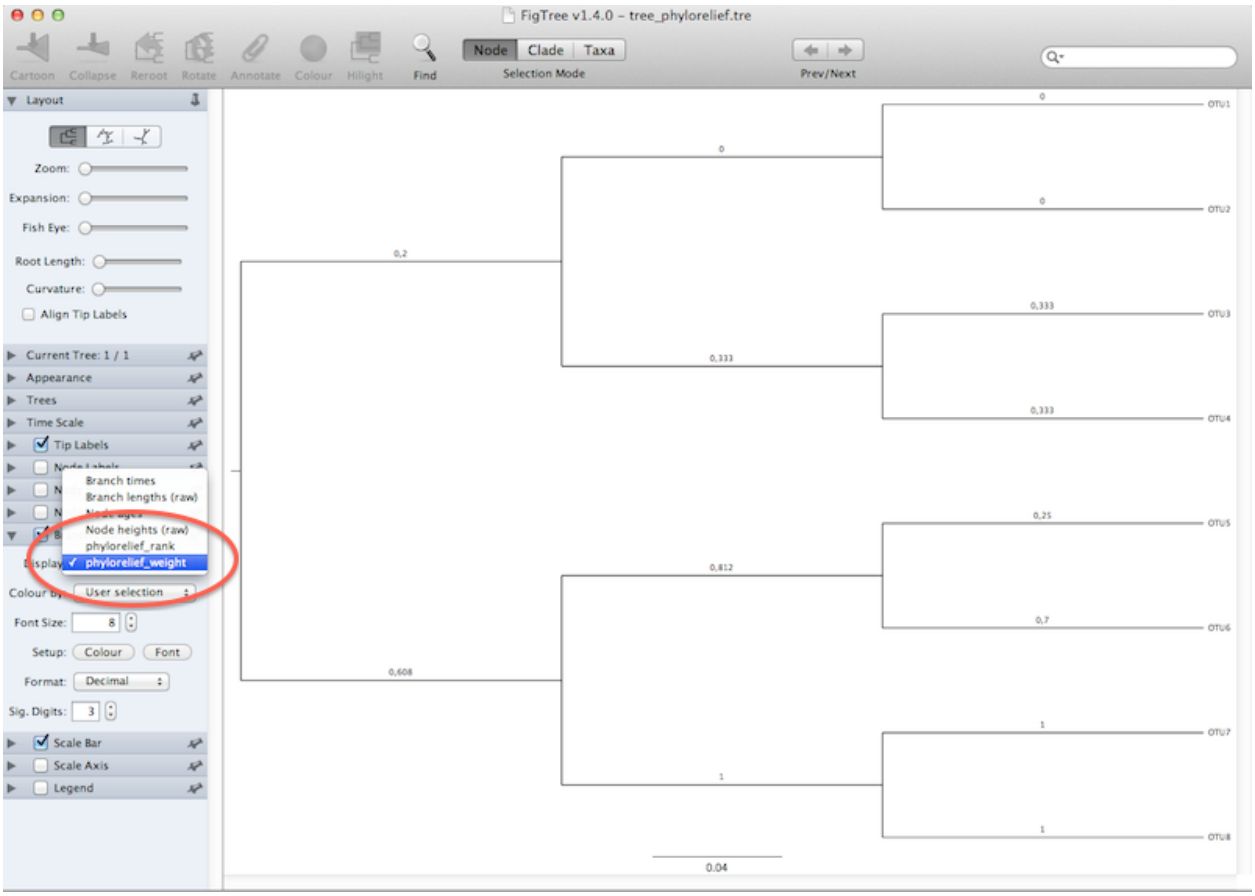
Two files, `out_phylorelief.txt` and `tree_phylorelief.tre` will be generated. `out_phylorelief.txt` contains the clade/OTU ranking:

```
phylorelief_weight        kw_stats        kw_p        kw_padj        OTUs
0.9762        3.0000        0.0833        0.1388        OTU7,OTU8
0.7976        3.0000        0.0833        0.1388        OTU5,OTU6
0.2961        3.0000        0.0833        0.1388        OTU3,OTU4
0.0703        1.0000        0.3173        0.3173        OTU2
0.0096        1.0000        0.3173        0.3173        OTU1
```

`tree_phylorelief.tre` is the nexus annotated file:

```
#NEXUS



BEGIN TAXA;

    DIMENSIONS NTAX=8;
    TAXLABELS
        OTU1
        OTU2
        OTU3
        OTU4
        OTU5
        OTU6
        OTU7
        OTU8
  ;
END;

BEGIN TREES;

    TREE 0 = [&R] (((OTU1:0.1[&phylorelief_weight=0.00960061443932,phylorelief_rank=5,kw_stat=1.0,kw_

END;
```

Now you can navigate the annotated tree with an external progam such as FigTree:


FigTree v1.4.0 – tree_phylorelief.tre
Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
Align Tip Labels
Current Tree: 1 / 1
Appearance
Trees
Time Scale
Tip Labels
Branch times
Branch lengths (raw)
Node heights (raw)
phylorelief_rank
phylorelief_weight
User selection
Font Size:
8
Setup:
Colour
Font
Format:
Decimal
Sig. Digits:
3
Scale Bar
Scale Axis
Legend
0,2
0
0
0,333
0,333
0,333
0,608
0,812
0,25
0,7
1
1
1
OTU1
OTU2
OTU3
OTU4
OTU5
OTU6
OTU7
OTU8
0.04

CHAPTER

FOUR

ALGORITHM DETAILS

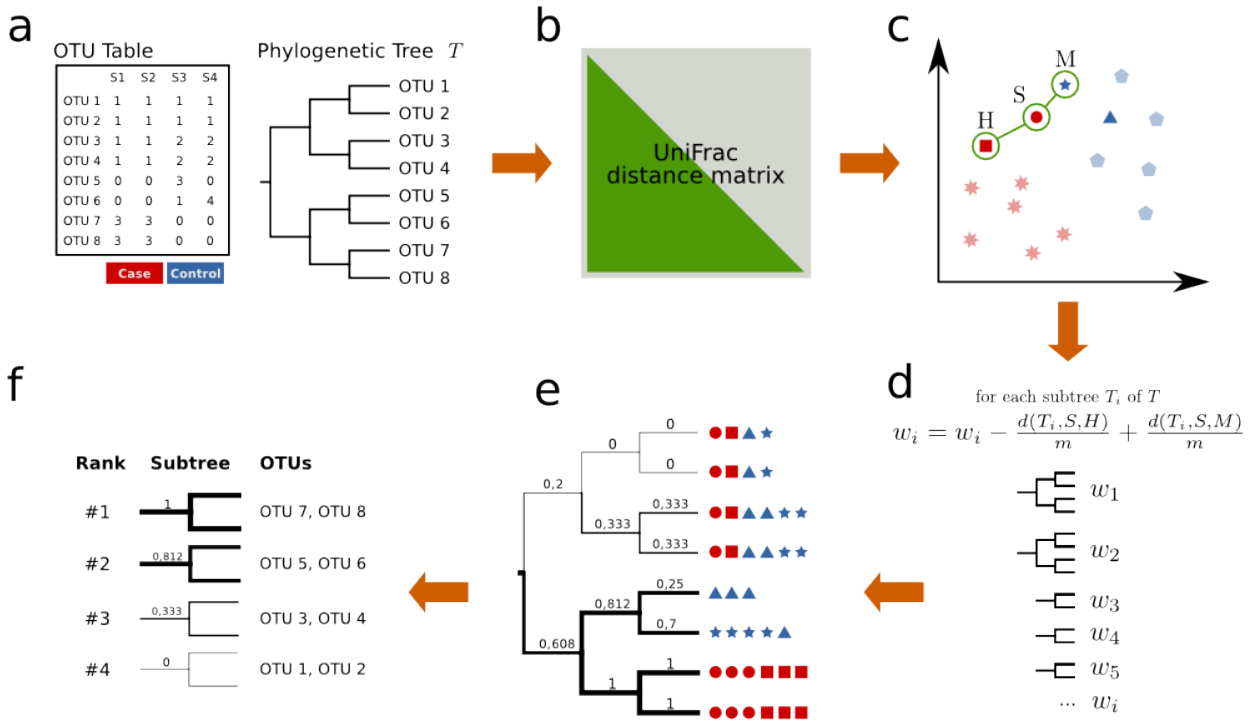


Figure 4.1: a) Preliminary analysis. From the sequences of the marker genomic loci selected by the experimental design, an OTU table and a phylogenetic tree of the representative sequences of the OTUs is computed. b) Next, the matrix of the distances between the samples must be computed using a phylogenetic measure of β-diversity, such as weighted or unweighted UniFrac must be provided.; c) The PhyloRelief strategy. Once one sample S has been randomly selected, the nearest hit H, i.e. the nearest sample of the same class, and the nearest miss M, i.e. the nearest sample of different class according to distance matrix D^S are identified. d) The update function. For each branch T_i the weight $W[T_i]$ is updated by summing the value $d(T_i, S, H)/m$ and subtracting $d(T_i, S, M)/m$. The function $d(T_i, A, B)$ is computed by summing the UniFrac distance between the sample A and B restricted to the subtree T_i. e) Correlation of the weights and definition of the clades. The weights of the each clades propagate to the parents, where it is either reinforced if coalescing with a clade sharing similar unbalance between the classes, or is diluted if coalescing with a clade with no or contrasting unbalance. This allows an iterative procedure leading to the the unambiguous identification of a set of uncorrelated clades. f) Output. The algorithm provides a list of clades of the phylogenetic tree ranked according to their contribution to the separation of the classes of samples.

CHAPTER

FIVE

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*