FONDAZIONE EDMUND MACH
CENTRO RICERCA e INNOVAZIONE

SYMBIOMES: SYSTEMS BIOLOGY OF HOST-MICROBIOME INTERACTIONS
5-10 JUNE 2015, PULTUSK. POLAND

PhyloRelief

phylogenetic-based framework for OTU weighting and selection

Davide Albanese, Carlotta De Filippo, Duccio Cavalieri, Claudio Donati*
Fondazione Edmund Mach, San Michele all'Adige (TN), Italy

Motivation

Metagenomics is revolutionizing our understanding of microbial communities, showing that their structure and composition have profound effects on the ecosystem and in a variety of health and disease conditions. In **case/control studies**, a common task is to estimate the **relevance** *w* (e.g. using univariate tests, as t-test) of each Operational Taxonomic Unit (OTU) and/or their **best predictive subset** applying ML algorithms, like Random Forest, for classification purposes (e.g. medical diagnosis and forensics identification) (Knights et al. 2010). Current statistical and learning approaches take as input alternatively:

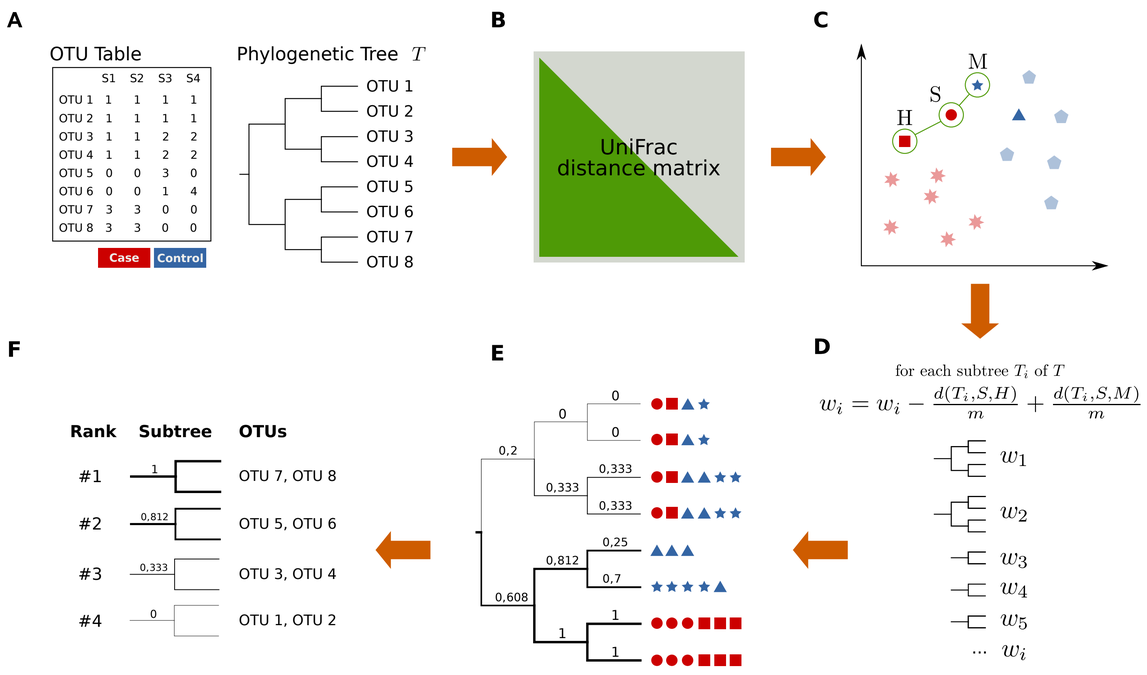
(i) A sample-by-OTU abundance matrix;
(ii) A sample-by-taxa (e.g. OTU matrix merged at the genus level) abundance matrix after a taxonomic classification.

In the former case *(i)* the analysis do not account for the **different degrees of similarity between sequences**, and in latter case *(ii)* the taxonomic classification often does not allow an **adequate description of the structure** of the microbiota.

Our solution:

integrating the phylogenetic information into the OTU relevance estimation process, without relying on pre-defined taxonomic categories

How PhyloRelief works



A) Inputs: an OTU table and a phylogenetic tree of the representative sequences

B) PhyloeRelief computes the matrix of the distances between the samples using a phylogenetic measure of β-diversity, such as weighted or unweighted UniFrac.

C) PhyloRelief randomly selects one sample *S* and identifies its nearest hit *H*, i.e. the nearest sample of the same class, and the nearest miss *M*, i.e. the nearest sample of the different class according to distance matrix.

D) Relief (Kira et al. 1992) strategy: for each subtree T_i PhyloRelief updates the weight w_i by summing d(T_i,*S*,*M*)/*m* and subtracting d(T_i,*S*,*H*)/*m*. The function d(T_i,*A*,*B*)/*m* is computed by summing the UniFrac distance between the sample *A* and *B* restricted to the subtree T_i and *m* is the number of samples:

unweighted PhyloRelief

$$d(\mathcal{T}_i, A, B) = \frac{\sum_{\mathcal{B}_q \in \{\mathcal{T}_i\}} b_q |\Theta_q^A - \Theta_q^B|}{\sum_{\mathcal{B}_q \in \{\mathcal{T}_i\}} b_q}$$

weighted PhyloRelief

$$d(\mathcal{T}_i, A, B) = \frac{\sum_{\mathcal{B}_q \in \{\mathcal{T}_i\}} b_q |p_q^A - p_q^B|}{\sum_{\mathcal{B}_q \in \{\mathcal{T}_i\}} b_q (p_q^A + p_q^B)}$$

E) The weights of each clade propagate to the parents: 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 unambiguous identification of a set of uncorrelated clades.

F) PhyloRelief provides a list of clades of the phylogenetic tree ranked according to their contribution to the separation of the classes of samples.

Analogously to the Relief-F (Kononenko, 1994) algorithm, PhyloRelief can work with multi-class classification problems. Moreover, a more robust form is also available: in **C)** for each sample *S*, *k* nearest neighbors from the same class H_i and *k*- nearest misses $M_i(C)$ are identified.

Predictive classification pipeline

We compared the predictive performances of PhyloRelief coupled with the Random Forest (RF) classifier (PhyloRelief +RF) to LEfSe + RF, MetaPhyl (without feature selection) and RF alone used both as classifier and feature selection method (RF + RF).

- Predictive pipeline based on a stratified 10x random subsampling cross validation (CV);
- To avoid selection bias effects, the OTU selection procedure was included in the CV loop;
- For each training set, the number of ranked features n_0 that provides the smallest average error is found by a nested 10x random subsampling CV. Later, the features are ranked using the entire training set and the model is trained using the top ranked n_0 features;
- Publicly available datasets recently used as benchmark in comparative evaluations of classification:
  * Costello et al. 2009 Body Habitats (CBH) dataset: forehead (FH) vs. external nose (EN)
  * Costello et al. 2009 Body Habitats (CBH) dataset: volar forearm (VF) vs. popliteal fossa (PF)
  * Papa et al. 2012 IBD dataset from fecal samples: IBD vs. healthy
  * Fierer et al. 2010 forensic skin (FS) dataset: subject identification (3 classes)
  * Fierer et al. 2010 forensic skin (FS) dataset: subject/hand identification (6 classes).

		FH vs. EN (CBH)	VF vs. PF (CBH)	IBD	FS subject (C = 3)	FS subject/hand (C = 6)
PhyloRelief W + RF	k = 2	0.214 0.103 (4)	0.655 0.045 (800)	-0.011 0.060 (40)	**1 0 (700)**	0.678 0.028 (900)
	k = 3	0.158 0.060 (4)	0.718 0.033 (800)	0.079 0.090 (40)	**1 0 (700)**	0.666 0.027 (800)
	k = 4	**0.220 0.073 (4)**	0.685 0.065 (800)	0.074 0.067 (40)	**1 0 (700)**	**0.684 0.026 (900)**
PhyloRelief U + RF	k = 2	-0.042 0.087 (4)	0.565 0.077 (800)	0.165 0.057 (40)	**1 0 (700)**	0.655 0.024 (900)
	k = 3	0.112 0.095 (4)	0.539 0.080 (800)	0.213 0.074 (40)	0.994 0.006 (700)	0.640 0.020 (800)
	k = 4	0.066 0.089 (4)	0.599 0.050 (800)	0.121 0.078 (40)	0.994 0.006 (700)	0.653 0.017 (900)
LEfSe + RF	OTU	-0.039 0.061 (19)	**0.836 0.040 (100)**	0.083 0.057 (81)	**1 0 (181)**	0.628 0.022 (59)
	Taxa	0.044 0.059 (4)	0.833 0.035 (50)	**0.238 0.065 (20)**	0.983 0.008 (85)	0.517 0.034 (101)
RF	FS	0.108 0.099 (1)	0.784 0.074 (40)	0.142 0.059 (7)	1.0 0.0 (200)	0.670 0.026 (30)
	No FS	-0.021 0.021 (-)	0.659 0.060 (-)	0.0 0.0 (-)	1.0 0.0 (-)	0.667 0.026 (-)
MetaPhyl	No FS	0.170 0.106 (-)	0.831 0.048 (-)	0.229 0.085 (-)	0.950 0.022 (-)	0.672 0.036 (-)

Classification accuracy in terms of average K-category correlation coefficient (KCCC) using weighted and unweighted PhyloRelief, LEfSe using OTUs and classified taxa, RF and MetaPhyl. In parentheses, the number of selected OTUs (best model).

Case study: gut microbiota of Red Colobus monkeys

Seven social groups inhabiting two forests in the Udzungwa Mountains of Tanzania (Barelli et al. submitted): Magombera (disturbed) vs. Mwanihana (undisturbed).

The most relevant clades selected by PhyloRelief (Kruskal-Wallis test, P<0.01) highlight that, beside more evident differences, there is a general rearrangement of the taxa within the *Bacteroidales* and *Clostridiales* order, resulting in a lower diversity of the microbiota of the Magombera individuals.

Software

PhyloRelief is an **Open Source** project and it is implemented in Python. Requirements: NumPy/SciPy, Pandas, DendroPy and Statsmodels libraries. PhyloRelief software and the predictive classification pipeline are available at:

http://compmetagen.github.io/phylorelief

Example of command line usage:

```
$ phylorelief otu_table.txt tree.tre sample_data.txt Status -k 2 -u weighted
```

- `otu_table.txt`: TAB-delimited sample-by-OTU abundance matrix
- `tree.tre`: Rooted phylogenetic tree
- `sample_data.txt Status`: TAB-delimited metadata file and target column
- `-k 2`: number of nearest neighbors
- `-u weighted`: weighted distance

Output: a clade ranking file and an annotated tree file in NEXUS format.

References

(Albanese et al. 2015) *Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting. PLOS Computational Biology, 2015.*
(Knights et al. 2010) *Supervised classification of human microbiota*. FEMS microbiology reviews, 2010.
(Kira et al. 1992) *The feature selection problem: Traditional methods and new algorithm*. In Proceedings of AAAI'92, 1992.
(Kononenko, 1994) *Estimating attributes: Analysis and extensions of Relief*. In Machine Learning: ECML-94, 1994.
(Barelli et al. submitted) *Habitat fragmentation is associated to gut microbiota diversity of an endangered primate: implications for conservation*. Submitted.