Sourcepredict: Prediction/source tracking of metagenomic samples source using machine learning

3rd May 2019

Summary

SourcePredict (github.com/maxibor/sourcepredict) is a Python package to classify and predict the source of metagenomics sample given a training set.

The DNA shotgun sequencing of human, animal, and environmental samples opened up new doors to explore the diversity of life in these different environments, a field known as metagenomics (Hugenholtz and Tyson 2008).
One of the goals of metagenomics is to look at the composition of a sequencing sample with tools known as taxonomic classifiers. These taxonomic classifiers, such as Kraken (Wood and Salzberg 2014) for example, will compute the taxonomic composition in Operational Taxonomic Unit (OTU), from the DNA sequencing data.

When in most cases the origin of a metagenomic sample is known, it is sometimes part of the research question to infer and/or confirm its source.
Using samples of known sources, a training set can be established with the OTU sample composition as features, and the source of the sample as class labels.
With this training set, a machine learning algorithm can be trained to predict the source of unlabeled samples from their OTU taxonomic composition.

Here, I developed SourcePredict to perform the classification/prediction of unlabeled samples sources from their OTU taxonomic compositions.

Method

All samples are first normalized to correct for uneven sequencing depth using GMPR (default) (Chen et al. 2018). After normalization, Sourcepredict performs a two steps prediction.

Prediction of unknown sources proportion

The unknown sources proportion is the proportion of OTUs in the test sample which are not present in the training dataset.

Let S be a sample of size O with O OTUs from the test dataset D_{test}
Let n be the average number of samples per class in the training dataset.
Let U_n be the samples to add to the training dataset to account for the unknown source proportion in a test sample.

First a α proportion (default=0.1) of each o_i OTU (with $i \in [1, O]$) is added to the training dataset for each U_j samples (with $j \in [1, n]$), such as $U_j(o_i) = \alpha \times S_(o_i)$

The U_n samples are then merged as columns to the training dataset (D_{train}) to create a new training dataset denoted $D_{train\ unknown}$

To predict this unknown proportion, the dimension of the training dataset $D_{train\ unknown}$ (samples in columns, OTUs as rows) is first reduced to 20 with the scikit-learn (Pedregosa et al. 2011) implementation of the PCA.
This training dataset is further divided into three subsets: train (64%), test (20%), and validation (16%).
The scikit-learn implementation of K-Nearest-Neighbors (KNN) algorithm is then trained on the train subset, and the test accuracy is computed with the test subset.
The trained KNN model is then corrected for probability estimation of unknown proportion using the scikit-learn implementation of the Platt's scaling method (Platt and others 1999) with the validation subset. This procedure is repeated for each sample of the test dataset.

Prediction of known source proportion

First, only OTUs corresponding to the *species* taxonomic level are kept using ETE toolkit (Huerta-Cepas, Serra, and Bork 2016). A distance matrix is then computed on the merged training dataset D_{train} and test dataset D_{test} using the scikit-bio implementation of weighted Unifrac distance (default) (Lozupone et al. 2007).

The distance matrix is then embedded in two dimensions using the scikit-learn implementation of t-SNE (Maaten and Hinton 2008).

The 2-dimensional embedding is then split back to training and testing dataset.

The training dataset is further divided into three subsets: train (64%), test (20%), and validation (16%).
The scikit-learn implementation of K-Nearest-Neighbors (KNN) algorithm is then trained on the train subset, and the test accuracy is computed with the test subset.

The trained KNN model is then corrected for source proportion estimation using the scikit-learn implementation of the Platt's method with the validation subset.

Combining unknown and source proportion

For each sample, the predicted unknown proportion *p_unknown* is then combined with the predicted proportion of each of the C source class c of the training dataset such as:

$$\sum_{c=1}^{C} s_c + p_{unknown} = 1$$

with

$$s_c = s_{c\ predicted} \times p_{unknown}$$

CLI

The SourcePredict CLI is handled with ArgParse. A typical command to use SourcePredict is as simple as:

```
sourcepredict path/to/test_otu_table.csv
```

The documentation of CLI is available at sourcepredict.readthedocs.io

References

Chen, Li, James Reeve, Lujun Zhang, Shengbing Huang, Xuefeng Wang, and Jun Chen. 2018. "GMPR: A Robust Normalization Method for Zero-Inflated Count Data with Application to Microbiome Sequencing Data." *PeerJ* 6: e4600. https://doi.org/10.7717/peerj.4600.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data." *Molecular Biology and Evolution* 33 (6): 1635–8. https://doi.org/10.1093/molbev/msw046.

Hugenholtz, Philip, and Gene W Tyson. 2008. "Microbiology: Metagenomics." *Nature* 455 (7212): 481. https://doi.org/10.1038/455481a.

Lozupone, Catherine A, Micah Hamady, Scott T Kelley, and Rob Knight. 2007. "Quantitative and Qualitative Beta Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities." *Appl. Environ. Microbiol.* 73 (5): 1576–85. https://doi.org/10.1128/AEM.01996-06.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-Sne." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–30.

Platt, John, and others. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." *Advances in Large Margin Classifiers* 10 (3): 61–74.

Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3): R46. https://doi.org/10.1186/gb-2014-15-3-r46.