Sourcepredict: Prediction/source tracking of metagenomic sample sources using machine learning

15th May 2019

Summary

SourcePredict (github.com/maxibor/sourcepredict) is a Python Conda package to classify and predict the source of metagenomics sample given a reference dataset of known sources.

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 aspect of metagenomics is to look at the organism composition in a sequencing sample, with tools known as taxonomic classifiers. These taxonomic classifiers, such as Kraken (Wood and Salzberg 2014) for example, will compute the organism taxonomic composition, from the DNA sequencing data.

When in most cases the origin (source) 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 reference dataset can be established with the samples taxonomic composition (the organisms identified in the sample) as features, and the source of the sample as class labels. With this reference dataset, a machine learning algorithm can be trained to predict the source of unlabeled samples from their taxonomic composition.
Compared to SourceTracker (Knights et al. 2011), which uses gibbs sampling, Sourcepredict uses dimension reduction algorithms, followed by K-Nearest-Neighbors (KNN) classification.

Here, I present SourcePredict for the classification/prediction of unlabeled sample sources from their 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: first a prediction of the proportion of unknown sources, i.e. not represented in the reference dataset. Then a prediction of the proportion of each known source of the reference dataset in the test samples.

Organism are represented by their taxonomic identifiers (TAXID).

Prediction of unknown sources proportion

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

To compute U_n, a α proportion (default = 0.1) of each o_i organism (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 reference dataset (D_{ref}) to create a new reference dataset denoted $D_{ref\ unknown}$

To predict this unknown proportion, the dimension of the reference dataset $D_{ref\ unknown}$ (samples in columns, organisms as rows) is first reduced to 20 with the scikit-learn (Pedregosa et al. 2011) implementation of PCA.
This reference dataset is then divided into three subsets: $D_{train\ unknown}$ (64%), $D_{test\ unknown}$ (20%), and $D_{validation\ unknown}$(16%).

The scikit-learn implementation of KNN algorithm is then trained on $D_{train\ unknown}$, and the test accuracy is computed with $D_{test\ unknown}$.
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 $D_{validation\ unknown}$. This procedure is repeated for each sample of the test dataset.

The proportion of unknown $p_{unknown}$ sources in each sample is then computed using the trained and corrected KNN model.

Prediction of known source proportion

First, only organism TAXID 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_{ref} and test dataset D_{sink} 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 $D_{ref\ tsne}$ and testing dataset $D_{sink\ tsne}$.

The training dataset $D_{ref\ tsne}$ is further divided into three subsets: $D_{train\ tsne}$ (64%), $D_{test\ tsne}$ (20%), and $D_{validation\ tsne}$ (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 $D_{test\ tsne}$.
The trained KNN model is then corrected for source proportion estimation using the scikit-learn implementation of the Platt's method with $D_{validation\ tsne}$.

The proportion of each source p_c sources in each sample is then computed using the trained and corrected KNN model.

Combining unknown and source proportion

Finally, for each sample, the predicted unknown proportion $p_{unknown}$ is then combined with the predicted proportion p_c of each of the C source class c of the training dataset such as:

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

with

$$s_c = p_c \times p_{unknown}$$

Finally, a summary table is created to gather the estimated sources proportions.

Acknowledgements

Thanks to Dr. Alexander Herbig for proofreading this manuscript.

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.

Knights, Dan, Justin Kuczynski, Emily S Charlson, Jesse Zaneveld, Michael C Mozer, Ronald G Collman, Frederic D Bushman, Rob Knight, and Scott T Kelley. 2011. "Bayesian Community-Wide Culture-Independent Microbial Source Tracking." *Nature Methods* 8 (9): 761.

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.