The Journal of Open Source Software

# Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification

**Maxime Borry**[1]

**1** Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Germany

**DOI:**

**Software**

- Review
- Repository
- Archive

**Submitted:**
**Published:**



## Summary

SourcePredict (github.com/maxibor/sourcepredict) is a Python Conda package to classify and predict the origin of metagenomic samples, given a reference dataset of known origins, a problem also known as source tracking.

DNA shotgun sequencing of human, animal, and environmental samples has opened up new doors to explore the diversity of life in these different environments, a field known as metagenomics (Hugenholtz & Tyson, 2008).
One aspect of metagenomics is investigating the community composition of organisms within a sequencing sample with tools known as taxonomic classifiers. These taxonomic classifiers, such as Kraken (Wood & Salzberg, 2014), will compute the organism taxonomic composition from the DNA sequencing data.

In cases where the origin of a metagenomic sample, its source, is unknown, it is often part of the research question to predict and/or confirm the source. Using samples of known sources, a reference dataset can be established with the taxonomic composition of the samples, *i.e.* the organisms identified in the samples as features, and the sources of the samples as class labels. With this reference dataset, a machine learning algorithm can be trained to predict the source of unknown samples (sinks) from their taxonomic composition.
Other tools used to perform the prediction of a sample source already exist, such as SourceTracker (Knights et al., 2011), which employs Gibbs sampling. However, with Sourcepredict using a dimension reduction algorithm, followed by K-Nearest-Neighbors (KNN) classification, the interpretation of the results is made more straightforward thanks to the embedding of the samples in a human observable low dimensional space.

## Method

Starting with a numerical organism count matrix (samples as columns, organisms as rows, obtained by a taxonomic classifier) of merged references and sinks datasets, samples are first normalized relative to each other, to correct for uneven sequencing depth using the GMPR method (default) (Chen et al., 2018). After normalization, Sourcepredict performs a two-step prediction algorithm. First, it predicts the proportion of unknown sources, *i.e.* which are not represented in the reference dataset. Second it predicts the proportion of each known source of the reference dataset in the sink samples.

Organisms are represented by their taxonomic identifiers (TAXID).

### Prediction of unknown sources proportion

Let $S_i \in \{S_1, .., S_n\}$ be a sample from the normalized sinks dataset $D_{sink}$, ${o_j}^i \in \{{o_1}^i, .., {o_{n_o}}^i\}$ be an organism in $S_i$, and ${n_o}^i$ be the total number of organisms in $S_i$, with ${o_j}^i \in \mathbb{Z}+$.
Let $m$ be the mean number of samples per class in the reference dataset, such that $m = \frac{1}{O}\sum_{i=1}^{O} S_i$.
For each $S_i$ sample, I define $||m||$ estimated samples $U_k^{S_i} \in \{U_1^{S_i}, .., U_{||m||}^{S_i}\}$ to add to the reference dataset to account for the unknown source proportion in a test sample.

Separately for each $S_i$, a proportion denoted $\alpha \in [0, 1]$ (default = 0.1) of each of the ${o_j}^i$ organism of $S_i$ is added to each $U_k^{S_i}$ samples such that $U_k^{S_i}({o_j}^i) = \alpha \cdot x_{i\ j}$ , where $x_{i\ j}$ is sampled from a Gaussian distribution $\mathcal{N}\big(S_i({o_j}^i), 0.01\big)$.

The $||m||$ $U_k^{S_i}$ samples are then added to the reference dataset $D_{ref}$, and labeled as *unknown*, to create a new reference dataset denoted ${}^{unk}D_{ref}$.

To predict the proportion of unknown sources, a Bray-Curtis (Bray & Curtis, 1957) pairwise dissimilarity matrix of all $S_i$ and $U_k^{S_i}$ samples is computed using scikit-bio. This distance matrix is then embedded in two dimensions (default) with the scikit-bio implementation of PCoA.
This sample embedding is divided into three subsets: ${}^{unk}D_{train}$ (64%), ${}^{unk}D_{test}$ (20%), and ${}^{unk}D_{validation}$(16%).

The scikit-learn implementation of KNN algorithm is then trained on ${}^{unk}D_{train}$, and the training accuracy is computed with ${}^{unk}D_{test}$.
This trained KNN model is then corrected for probability estimation of the unknown proportion using the scikit-learn implementation of Platt's scaling method (Platt & others, 1999) with ${}^{unk}D_{validation}$.

The proportion of unknown sources in $S_i$, $p_u \in [0, 1]$ is then estimated using this trained and corrected KNN model.

Ultimately, this process is repeated independantly for each sink sample $S_i$ of $D_{sink}$.

### Prediction of known source proportion

First, only organism TAXIDs corresponding to the species taxonomic level are retained using the ETE toolkit (Huerta-Cepas, Serra, & Bork, 2016). A weighted Unifrac (default) (Lozupone, Hamady, Kelley, & Knight, 2007) pairwise distance matrix is then computed on the merged and normalized training dataset $D_{ref}$ and test dataset $D_{sink}$ with scikit-bio.

This distance matrix is then embedded in two dimensions (default) using the scikit-learn implementation of t-SNE (Maaten & Hinton, 2008).

The 2-dimensional embedding is then split back to training ${}^{tsne}D_{ref}$ and testing dataset ${}^{tsne}D_{sink}$.

The training dataset ${}^{tsne}D_{ref}$ is further divided into three subsets: ${}^{tsne}D_{train}$ (64%), ${}^{tsne}D_{test}$ (20%), and ${}^{tsne}D_{validation}$ (16%).
The KNN algorithm is then trained on the train subset, with a five (default) cross validation to look for the optimum number of K-neighbors. The training accuracy is then computed with ${}^{tsne}D_{test}$. Finally, this second trained KNN model is also corrected for source proportion estimation using the scikit-learn implementation of the Platt's method with ${}^{tsne}D_{validation}$.

The proportion $p_{c_s} \in [0, 1]$ of each of the $n_s$ sources $c_s \in \{c_1,\ ..,\ c_{n_s}\}$ in each sample $S_i$ is then estimated using this second trained and corrected KNN model.

### Combining unknown and source proportion

Then for each sample $S_i$ of the test dataset $D_{sink}$, the predicted unknown proportion $p_u$ is then combined with the predicted proportion $p_{c_s}$ for each of the $n_s$ sources $c_s$ of the training dataset such that $\sum_{c_s=1}^{n_s} s_c + p_u = 1$ where $s_c = p_{c_s} \cdot p_u$.

Finally, a summary table gathering the estimated sources proportions is returned as a `csv` file, as well as the t-SNE embedding sample coordinates.

## Acknowledgements

Thanks to Dr. Christina Warinner, Dr. Alexander Herbig, Dr. AB Rohrlach, and Alexander Hübner for their valuable comments and for proofreading this manuscript. This work was funded by the Max Planck Society and the Deutsche Forschungsgemeinschaft, project code: EXC 2051 #390713860.

## References

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, *27*(4), 325–349. doi:10.2307/1942268

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., & Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, *6*, e4600. doi:10.7717/peerj.4600

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, *33*(6), 1635–1638. doi:10.1093/molbev/msw046

Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, *455*(7212), 481. doi:10.1038/455481a

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., et al. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, *8*(9), 761. doi:10.1038/nmeth.1650

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, *73*(5), 1576–1585. doi:10.1128/AEM.01996-06

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Platt, J., & others. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61–74.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, *15*(3), R46. doi:10.1186/gb-2014-15-3-r46