The Journal of Open Source Software

Sourcepredict: Prediction of metagenomic sample sources using machine learning algorithms

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 source of metagenomics sample given a reference dataset of known sources, a problem also known as source tracking.

DNA shotgun sequencing of human, animal, and environmental sample 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 to investigate the community composition of organisms within a sequencing sample with tools known as taxonomic classifiers. These taxonomic classifiers, such as for example 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 this source. Using samples of known sources, a reference dataset can be established with the samples taxonomic composition, i.e. 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 unknown samples (sinks) from their taxonomic composition.
Other tools to perform the prediction of a sample source exist, such as SourceTracker (Knights et al., 2011), which uses 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: 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 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 of size O organisms o_j from the normalized sinks dataset D_{sink}, with $o_j \in \mathbb{Z}+$, and $j \in [1, O]$.
Let m be the mean number of samples per class in the reference dataset, such as $m = \frac{1}{O}\sum_{i=1}^{O} S_i$.

I define $|m|$ estimated samples U_k to add to the reference dataset to account for the unknown source proportion in a test sample, with $k \in \{1, .., |m|\}$.

To compute each U_k, a α proportion ($\alpha \in [0, 1]$, default $= 0.1$) of each o_j organism is added for each U_k samples of the reference dataset, such that $U_k(o_j) = \alpha \cdot x_{i\ j}$, where $x_{i\ j}$ is sampled from the Gaussian distribution $\mathcal{N}\big(\mu = S_i(o_j), \sigma = 0.1\big)$.

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

To predict the proportion of unknown sources, a distance matrix of the samples is computed using the scikit-bio implementation of the Bray-Curtis dissimilarity (Bray & Curtis, 1957). 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: $D_{train\ u}$ (64%), $D_{test\ u}$ (20%), and $D_{validation\ u}$(16%).

The scikit-learn implementation of KNN algorithm is then trained on $D_{train\ u}$, and the test accuracy is computed with \$D_{test u} \$.
This trained KNN model is then corrected for probability estimation of unknown proportion using the scikit-learn implementation of the Platt's scaling method (Platt & others, 1999) with $D_{validation\ u}$. This procedure is repeated for each S_i sample of the test dataset D_{sink}.

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

Prediction of known source proportion

First, only organism TAXID corresponding to the *species* taxonomic level are kept using ETE toolkit (Huerta-Cepas, Serra, & 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, Hamady, Kelley, & Knight, 2007).

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

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

The training dataset $D_{ref\ tsne}$ is further divided into three subsets: $D_{train\ t}$ (64%), $D_{test\ t}$ (20%), and $D_{validation\ t}$ (16%).
The KNN algorithm is then trained on the train subset, and the test accuracy is computed with $D_{test\ t}$.
This trained KNN model is then corrected for source proportion estimation using the scikit-learn implementation of the Platt's method with $D_{validation\ t}$.

p_c is then estimated using this trained and corrected KNN model, where $p_c \in [0, 1]$ is the proportion of each of source c in each sample S_i.

Combining unknown and source proportion

Finally, 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 for each of the C sources c of the training dataset such that $\sum_{c=1}^{C} s_c + p_u = 1$ where $s_c = p_c \cdot p_u$.

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

Acknowledgements

Thanks to Dr. Christina Warinner, Dr. Alexander Herbig, Dr. Adam Ben Rohrlach, and Alexander Hübner for their valuable comments and for proofreading this manuscript. This work was funded by the Max Planck Society.

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