Published in final edited form as:
Nat Methods. 2011 March ; 8(3): 191–192. doi:10.1038/nmeth0311-191.

Taxonomic metagenome sequence assignment with structured output models

K.R. Patil[1], **P. Haider**[3], **P.B. Pope**[4], **P.J. Turnbaugh**[5], **M. Morrison**[4], **T. Scheffer**[3], and **A.C. McHardy**[1,2,*]

[1]Max-Planck Research Group for Computational Genomics & Epidemiology, Max-Planck Institute for Informatics, Universitätscampus E1 4, 66123 Saarbrücken, Germany

[2]Department for Algorithmic Bioinformatics, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany

[3]University of Potsdam, Department of Computer Science, 14492 Potsdam, Germany

[4]CSIRO Livestock Industries, Queensland Bioscience Precinct, St Lucia, 4069 Australia

[5]Harvard FAS Center for Systems Biology, Northwest Lab Building, 52 Oxford Street, 435.40, Cambridge MA 02138, USA

To the editor

Computational inference of the taxonomic origin of sequence fragments is an essential step in metagenome analysis[1]. Assignment of fragments to individual populations or corresponding higher-level evolutionary clades can be performed with either homology-, sequence similarity- or sequence composition-based methods[2]. It is a challenging task, because for the majority of uncultured micro-organisms reference sequence is unavailable and large amounts of data have to be processed. With this in mind, we introduce PhyloPythiaS, a fast and accurate sequence compositional classifier based on the structured output paradigm[3].

We evaluated PhyloPythiaS on simulated and real metagenome data in comparison to four other methods; PhyloPythia[4], MEGAN[5], Phymm and PhymmBL[6]. PhyloPhythiaS performed particularly well for taxonomic assignment of populations from novel genera, order or higher-level clades, when limited amounts of reference data were available. Accurate assignments could be performed based on 100 kb of training data for a sample population. We observed this for simulated data (Fig. 1a, Supplementary Fig. 1 and Table 1) and a predominant population of a novel family of the order of Aeromonadales from the Australian Tammar wallaby gut (Fig. 1b, Supplementary Tables 5 and 9). In this scenario, alignment-based methods performed poorly. If closely related genomes were available the performance of all methods became more similar, with a slight advantage for alignment-based approaches. This is observed for simulated data and the predominant genera of two human gut metagenomes (Supplementary Tables 1.A and 11-13).

PhyloPythiaS also performed well in fragment assignment of ‘known unknowns’, i.e. for organisms of taxonomic clades with no available sequence. Here, we observed less ‘overbinning’, meaning assignments to correct higher-level, but incorrect low-level clades, than for PhymmBL (Supplementary Tables 5 and 8). For short fragments of ‘known

[*]Corresponding author; alice.mchardy@uni-duesseldorf.de.

unknowns', all methods showed comparably low assignment accuracy, with MEGAN performing best (Supplementary Fig. 2 and Table 2).

Empirical analysis of execution times determined that PhyloPythiaS requires 0.08-0.1 seconds for the assignment of 0.1-10 kb fragments (Fig. 1c). This corresponds to a three- to 46-fold and five- to 68-fold improvement in comparison to MEGAN and PhymmBL, respectively (Fig. 1c). For characterization of a 13 MB assembled metagenome sample, PhyloPythiaS showed 22-fold, 85-fold and 106-fold speed increase in comparison to PhyloPythia, MEGAN and PhymmBL, respectively (Supplementary Table 14). As PhyloPythiaS models require only a subsample of the reference data for accurate assignment, in future, training times will not necessarily be impacted by increases of sequence data, contrary to alignment-based approaches.

PhyloPythiaS employs an ensemble of linear models whose parameters are identified using the paradigm of support vector machines (SVM) with structured output spaces to represent composition-based clade specifics of the taxonomic hierarchy, instead of an ensemble of multi-class SVMs for different taxonomic ranks and fragment lengths, as our previously described method PhyloPythia (Supplementary notes). It exhibits considerable gains in learning and prediction times, while performing similarly by several independent measures on two real-world metagenome data sets (Supplementary Tables 5-7, 9 and 11-14). PhyloPythiaS is freely available for academic use.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank T. Joachims for the SVMstruct implementation, L. Steinbrück and L. Feuerbach for software testing. K.R.P and A.C.M. were funded by the Max-Planck society and Heinrich-Heine University Düsseldorf. T.S. gratefully acknowledges support from the German Science Foundation DFG. P.J.T. is funded by NIH P50 GM068763. The unpublished data for the WG-1 project arises from support provided by the OCE Science team of CSIRO Australia, in the form of an OCE postdoctoral fellowship (P.P.) and OCE Science Leader program (M.M.).

References

1. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. Microbiol Mol Biol Rev. 2008; 72(4):557–578. [PubMed: 19052320]
2. McHardy AC, Rigoutsos I. Curr Opin Microbiol. 2007; 10(5):499–503. [PubMed: 17933580]
3. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. J Mach Learn Res. 2005; 6:1453–1484.
4. McHardy AC, Garcia-Martin H, Tsirigos A, Hugenholtz P, R I. Nat Methods. 2007; 4(1):63–72. [PubMed: 17179938]
5. Huson DH, Auch AF, Qi J, Schuster SC. Genome Res. 2007; 17(3):377–386. [PubMed: 17255551]
6. Brady A, Salzberg SL. Nat Methods. 2009; 6(9):673–676. [PubMed: 19648916]

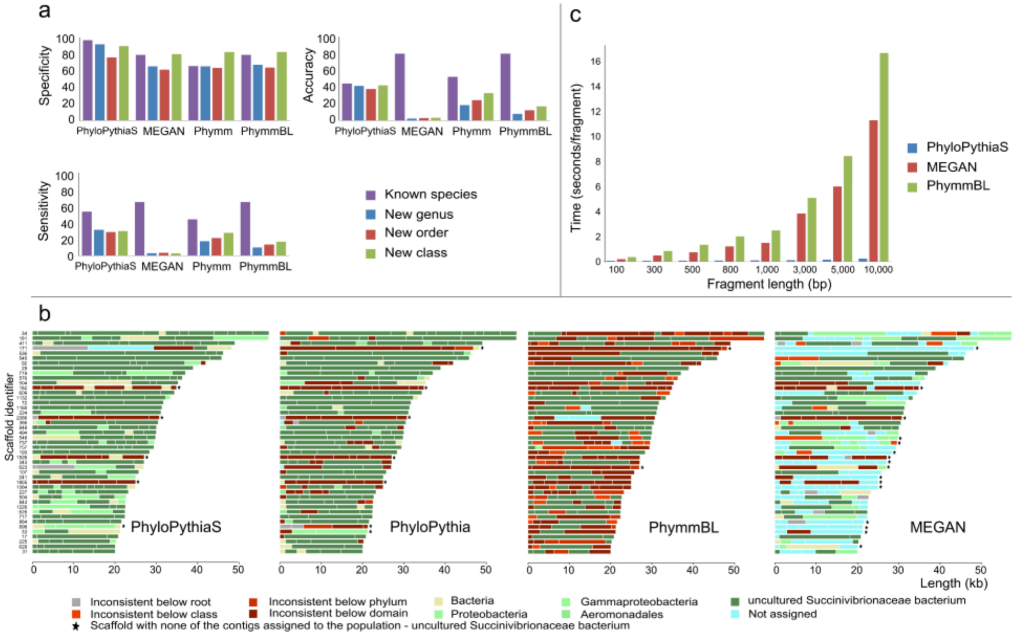

Figure 1.

Comparison of different taxonomic classification methods. (**a**) Average performance per contig for the simMC data set (average length 2332 bp) at genus rank in four different experiments. Each experiment reflects a different scenario in terms of available reference sequences from closely related organisms for the dominant strains. 'Known species' had complete genome sequences of the same species available as reference; in the other experiments sequences of the same clades at the respective rank were excluded, while retaining 100 kb of the dominant strains. (**b**) Scaffold-contig consistency for the WG-1 population (uncultured Succinivibrionaceae bacterium) of the Tammar wallaby gut metagenome. Contig coloring reflects taxonomic assignment consistency with respect to WG-1. Only scaffolds longer than 20 kb are shown. (**c**) Empirical execution time evaluated on a Linux machine with 3 GHz processor and 4 GB main memory. Results for MEGAN and PhymmBL were determined with a reference database of size 2.1 GB.