Supplementary Notes

Optimal oligonucleotide input space, sequence type, and other parameters

Both ‘raw’ genomic fragments and protein coding sequences were evaluated as suitable sequence sources for our technique. Both were found to carry a strong phylogenetic signal in sequence composition across all taxonomic ranks (**Supplementary Fig. 1**). Because of the highest similarity in terms of length of training items, the CDS model is best compared to the 1 kb model in this graph.

The clade-specific signal that is learned from genomic fragments is not identical to the signal that characterizes the genic regions, as the mutually decreasing performance of genomic fragments or genes with the other model demonstrates. This effect, however, is not very strong which indicates that a considerable part of the signal that is learned from both data types is frame-independent, possibly comprising patterns such as clade-specific restriction enzyme cleavage sites, regulatory and promoter elements, etc.

The decrease in performance for the CDS-trained model on the 1 kb genomic sequence fragments suggests that the well known 3-periodic preference for certain codons and codon combinations, which is generated by translational selection in protein encoding genes[1], also carries a very slight phylogenetic signal. In genomic fragments, this signal is weakened by the presence of non-coding regions and shifted into different frames.

We also performed an extensive search for the best oligonucleotide pattern space to subsequently use for phylogenetic classification. To this end, we evaluated patterns of lengths w with w ranging from 2 to 6 inclusive, allowing for w-$l \in \{0,1,2\}$ flexible positions in the composition template. For genomic sequence fragments, the lower ranking clades from the genus level to the class level can be discriminated best by 5-mers comprising contiguous nucleotides (**Supplementary Table 1**, **Supplementary Fig. 2**). For clades at the ranks of phylum and domain, more complex 6-mer patterns ((w=6, l=4) and (w=6, l=6), respectively) are needed to capture the characteristics of a joint ancestry best. For CDSs, starting with oligomers of length 3 or longer, in-frame patterns are generally more informative than not in-frame patterns. The most informative patterns at the level of the genus are in-frame hexamers comprising contiguous nucleotides. At the higher levels, less specific hexamers with two non-literal positions (=”wild-cards”) are more informative.

Relation of fragment length to classification accuracy

We evaluated the relation of classification accuracy for fragments of different lengths with classifiers trained on fragments of different lengths with 1, 3, 5, 10, 15, 50 kb fragments. For each length setting, a classifier was trained and used to classify fragments of differently sizes from the genomes of organisms unknown to the classifier. Assignment accuracy was evaluated with up to 100 fragments from each organism, depending on sequence availability.

The overall class-normalized classification accuracy (sensitivity) improves as the size of the training fragments decreases, but only up to the point where the tested fragments reach a similar size to the training fragments. For the levels from genus to phylum, the assignment specificity increases for longer classifiers trained with longer fragments across all fragment lengths tested. This convenient property results from the use of the second, binary classifier that more accurately rejects false positive assignments when trained with the less noisy sequence composition vectors of longer sequence fragments. The level of domain represents an exception in that the best classification accuracy is achieved with a classifier trained with fragments of similar size to the evaluated fragment. The reason for this different behavior is that there is no broadly defined ‘Other’ class at this level, to which items with unclear signal get assigned, and can then be retested with another model trained on shorter sequence fragments.