Analysis of barcode sequences by means of compression-based methods

M. La Rosa✉, A. Fiannaca, R. Rizzo, A. Urso
ICAR-CNR, National Research Council of IT, Palermo, Italy

Motivations

The key idea of DNA barcode initiative is to identify, for each group of species belonging to different kingdoms of life, a short DNA sequence that can act as a true taxon barcode. DNA barcode represents a valuable type of information that can be integrated with ecological, genetic, and morphological data in order to obtain a more consistent taxonomy. Recent studies have shown that, for the animal kingdom, the mitochondrial gene cytochrome c oxidase I (COI), about 650 bp long, can be used as a barcode sequence for identification and taxonomic purposes of animals. The analysis of DNA barcode sequences is carried out with well known bioinformatics techniques: for example the most common approach to create a phylogenetic tree for a group of species uses the multi-alignment of genetic sequences, the computation of a dissimilarity matrix, using one of the current available evolutionary distance model, and finally the building of a tree by means of hierarchical algorithms such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbor Joining (NJ). In the present work we aims at introducing the use of an alignment-free approach in order to make taxonomic analysis of barcode sequences. Our approach is based on the use of two compression-based versions of non-computable Universal Similarity Metric (USM) class of distances. This way we try to overcome some flaws of classic techniques, such as the time-consuming and parameter-dependent alignment procedure and the use of stochastic evolutionary distance models, that do not represent a distance metric.

Methods

Universal Similarity Metric represents a class of distance measures based on the non-computable Kolmogorov complexity. That means it needs some approximation in order to be used. USM is said to be "universal" because it can be applied for computing a distance matrix among input data belonging to very different application domains. In fact it has been used for the analysis of text, images, music. In bioinformatics, it has been applied for obtaining phylogenetic trees from complete mythocondrial genome of mammalian species. Our purpose is to justify the employ of USM also for the analysis of short DNA barcode sequences, showing USM is able to correctly extract taxonomic information among those kind of sequences. We, then, downloaded from Barcode of Life data System database (BOLD) 20 datasets of barcode sequences belonging to different animal species. For each dataset we computed dissimilarity matrices by means of two compression-based approximation of USM, namely Normalized Compression Distance (NCD) and its conditional compression version. In both cases we used GenCompres compressor, that is a dictionary-based compressor suited to work with DNA sequences. From those matrices we built, using UPGMA and NJ algorithms, phylogenetic trees of every dataset and compared them, in terms of topology preservation, with the trees obtained through Kimura 2-parameter evolutionary distance model.

Results

Experimental tests aim to evaluate the quality of phylogenetic reconstruction in terms of both topological similarity and differences in the relative branch length. As regard the tree similarity, we obtain good results with a percentage of similarity between evolutionary and compression-based tree greater than 82% and, for the most of datasets, between 90% and 100%. Lower results are for datasets having an high percentage of sequences with ambiguous bases. We detect the same trend for differences in the relative branch of trees, except that poorer results are reached by those datasets containing some COI-5P gene sequences longer than the other ones.